E-ISSN: 0974-4614
p-ISSN: 0972-0448

The Impact of PCA and t-SNE on the Predictive Accuracy of k-NN, Naive Bayes, and LDA: A Study Using the Legal Medicine Legal Medicine Dataset

P.Syamala Rao[1*], D Naga Malleswari[2], Koppula Srinivas Rao[3], Deepthi Kamidi[4],B.Sankara Babu[5], Kayam Saikumar[6]

[1]Assistant Professor, Department of IT, SRKR Engineering College,Email: peketi.shyam@gmail.com
[2]Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Green Fileds,Vaddeswaram, A.P. – 522302, Email:nagamalleswary@kluniversity.in
[3]Professor in Department of Computer Science and Engineering, MLR Institute of Technology, Hyderabad, Email: ksreenu2k@gmail.com
[4]Assistant Professor Department of Computer Science and Engineering Vignan Institute of Technology and Science, Email:deepthikamidi83@gmail.com
[5]Professor, Department of CSE, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Email: sankarababu.b@GRIET.ac.in
[6]Assistant Professor, department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Telangana, India, Email:saikumar.kayam@klh.edu.in
*Corresponding Author

Received: 17.08.2024 Revised: 18.09.2024 Accepted: 17.10.2024

ABSTRACT
Data is created in large quantities in today's digital world by many different industries, including healthcare, content creation, the internet, and business. Analysing this data to uncover relevant insights for decision-making is where machine learning (ML) algorithms come into play. However, not all features within these Legal Medicine Datasets contribute meaningfully to the construction of robust ML models. Some features may be irrelevant or have minimal impact on predictive performance. By removing these irrelevant features, the computational load on ML algorithms is alleviated. This research makes use of the open-source MNIST Legal Medicine Dataset to explore the relationship between dimensionality reduction techniques & various machine learning algorithms, such as k-Nearest Neighbours (k-NN), Naive Bayes, as well as Linear Discriminant Analysis (LDA). like t-SNE and PCA.The experimental results demonstrate the effectiveness of these ML algorithms in this context. Moreover, the study shows that incorporating PCA with ML algorithms enhances performance, especially when dealing with high-dimensional Legal Medicine Datasets.

Keywords: Dimensionality Reduction, KNN, ML, NB, PCA, LDA, t-SNE, SVM.

1.INTRODUCTION

The ability to decipher handwritten numbers is a major step forward in artificial intelligence a critical challenge with significant implications across various domains [1]. Significant strides have been achieved within the past 20 years, driven by extensive research into methodologies for handwritten digit recognition [2]. Despite these advances, machine learning algorithms often struggle with the curse of dimensionality in high-dimensional data [3]. As Legal Medicine Datasets expand, the accuracy of machine learning-based classification tends to decrease, highlighting the need for dimensionality reduction (DR) techniques to improve performance [4].

This paper explores the intersection of dimensionality reduction methods and machine learning algorithms, with a particular focus on their application to handwritten digit recognition. Handwritten digits, being a common form of communication, have garnered substantial global attention due to their widespread use [5]. To improve the precision of number recognition systems, we review the literature on dimensionality reduction methods and how they interact with machine learning frameworks. [6].

Through a review of key studies, including those by Md. Golam Sarowar et al., G. Thippa Reddy, Hany Yan, and others, we demonstrate the efficiency of non-dimensional methods for dimensionality reduction like PCA, LDA,and Singular Value Decomposition (SVD) in improving classifier performance [7]. We also address the potential drawbacks of using multiple DR techniques simultaneously, as highlighted by Gustavo et al., and emphasize the importance of choosing suitable DR techniques tailored to each Legal Medicine Dataset& classification job [8].

By synthesizing insights from various research efforts, this paper aims to advance the discussion on optimizing machine intelligence techniques for high-dimensional data analytics, with a particular emphasis on handwritten digit recognition.

2. LITERATURE SURVEY

As a crucial part of recognition approaches, handwriting digit recognition for the past two decades has been the focus of extensive study. An important factor influencing the success or failure of machine learning algorithms is the difficulty of dealing with data that has many dimensions. The reliability of machine learning-based categorisation may decline with an increase in the amount of measurement data, necessitating dimensionality reduction (DR) to enhance precision. Handwritten digit recognition, a critical task within computer vision, plays a crucial role in numerous applications. This process involves the complex integration of DR techniques and machine learning algorithms to effectively identify handwritten numerals, a task of growing importance due to its widespread use in daily life.

Extensive literature has explored various frameworks and mathematical approaches to digit recognition, with particular emphasis on improving accuracy and overall performance through the combination and development of different learning methodologies. A DR technique is essential when working with databases that contain a high dimensionality. Aiming to lower the dimensionality of initial data, these techniques, which is crucial for both learning frameworks and data mining. By minimizing the number of features, DR not only enhances classifier performance but also reduces computational complexity. There has been research on dimensionality reduction methods that are both linear and nonlinear.

The MNIST Legal Medicine Dataset, a benchmark in digit recognition, has been extensively studied, with various classifiers yielding unique accuracy results. For example, a PCA-based CNN with ACO achieved the highest accuracy of 80.84%.

Research by Adiwijaya et al. [1] compared PCA with SVM and LMBP algorithms within a cancer detection framework based on microarray data, warning against indiscriminate use of PCA.

Gustavo et al. [2] cautioned that combining multiple DR techniques could sometimes lead to worse outcomes than using a single method.

Pitoyo Hartono [3], working with the MNIST Legal Medicine Dataset, demonstrated the class accuracy achievable by embedding elegant data into its investigated numerous dimensionality reduction methods, including t-SNE, PCA, and NCA, in addition to low-dimensional representation utilising rRBF.

The study by Md. Golam Sarowar et al. (2020) [4] evaluates the performance of various various ML classifiers using feature extraction & dimensionality reduction methods. A presentation made at the IACC 9, this research provides a comparative analysis that highlights the impact of these techniques on classifier accuracy and computational efficiency. The findings underscore the significance of selecting appropriate dimensionality reduction methods to boost the efficiency of ML models generally, particularly in high-dimensional data scenarios.

Hany Yan et al. [5] highlighted the effectiveness of applying appropriate DR techniques prior to training to boost classification accuracy and reduce storage requirements, thus lowering the computational complexity involved in digit recognition.

G. Thippa Reddy et al. [6] observed that classifiers employing PCA outperformed those using LDA on the Cardiotocography (CTG) Legal Medicine Dataset, though they cautioned that DR methods must be carefully evaluated when applied to high-dimensional data such as text and images.

In order to improve handwritten digit recognition, Areej Alsaafin & Ashraf Elnagar (2017) [7] selected a subset of characteristics using feature selection methods. Featured in the AI Journal: Systems and Applications, their research emphasizes the importance of reducing feature dimensionality with the goal of enhancing computing efficiency & the precision of categorisation. The research demonstrates that by carefully selecting a minimal set of features, it is possible to achieve high recognition performance while minimizing the complexity of the machine learning model.

Mardani et al. (2020) [8] developed a multi-step process for estimating future CO_2 emissions by combining dimensionality reduction, clustering, and ML methods. This study demonstrates how dimensionality reduction can simplify large Legal Medicine Datasets; this, in turn, can improve the accuracy and computational efficiency of environmental impact assessment models when used in conjunction with clustering and other cutting-edge machine learning algorithms.

Rizgar R. Zebari et al. (2020) [9] included an extensive analysis of methods for reducing dimensionality in feature extraction and feature selection. Their analysis thoroughly investigates different approaches to data dimensionality reduction, highlighting the significance of these strategiesin enhancing the efficiency and performance of machine learning models. By offering insights into both feature selection and extraction approaches, the study serves as a valuable resource for understanding how to effectively manage high-dimensional Legal Medicine Datasets, thereby improving model accuracy and reducing computational complexity.

Drishti Beohar and Akhtar Rasool (2021) [10] studied the use of deep learning methods, particularly CNN and ANN, for recognising handwritten digits on the MNIST Legal Medicine Dataset. findings demonstrate the efficacy of cutting-edge neural networks in achieving high accuracy in digit recognition tasks. The study highlights the advantages of using CNNs over traditional methods, emphasizing their superior performance in handling the complexities of handwritten digit classification.
Tausif et al. [11] developed a lightweight CNN model specifically for the MNIST Legal Medicine Dataset, focusing on optimizing execution time. Additionally, compared different models using various machine learning classifier techniques on high-dimensional data.

3. Machine Learning Techniques

Different ML classification methods are

3.1 Naive Bayes algorithm

A straightforward and effective classification approach, Features must be independent for the Naive Bayes approach to work, as stated in Bayes' Theorem [12]. Its efficiency and effectiveness make it widely applicable across various domains, including medical diagnosis, spam filtering, and text categorization, particularly for large Legal Medicine Datasets. Naive Bayes is a probabilistic classification that uses a collection of features and Bayes' Theorem to make predictions about the likelihood of a class [13]. Here is the equation that states the theorem:

$$P(c|x) = \frac{P(x|c)*P(c)}{P(x)} \quad (1)$$

- $P(c|x)$represents the posterior probability of the given predictor (x, characteristics) for the class (c, target).
- $P(c)$ is the prior probability of the class.
- $P(x|c)$ is the likelihood, which is the probability of the predictor given the class.
- $P(x)$ is the prior probability of the predictor.

To streamline calculations, Assuming that, regardless of the class, all features are independent, which results in the following expansion:

$$P(c|x) = P(x1|c) \times P(x2|c) \times \ldots \times P(xn|c) \times P(c) \quad (2)$$

Here:

- x1, x2, …, xn are individual features within the feature set x.
- $P(xi|c)$ represents the likelihood of the i-th attribute, considering the category c

This independence assumption allows the algorithm to compute the posterior probability efficiently, even when working with data that has many dimensions [13]. The following step is to select the one that best fits the given set of features as the predicted classwith the highest posterior probability P(c|x) [14].

3.2 Support Vector Machine (SVM)

To accurately classify incoming data points, Support Vector Machines (SVM) seek out the optimal decision border that can effectively divide an n-dimensional space into several classes [15]. This boundary, called a hyperplane, is considered optimal because it maximizes the margin between the classes. SVM achieves this by identifying key data points, they are utilised to establish the hyperplane; these are called support vectors [16]. Support Vector Machines (SVMs) are a type of supervised learning algorithm that are optimised for tasks like this [17].

3.3 K-Nearest Neighbour (KNN) Algorithm

Supervised learning methods like K-Nearest Neighbour (KNN) [18] are useful for both regression and classification, albeit the former is where it really shines. K-Nearest Neighbours (KNN) is an instance-based or lazy learner algorithm that uses the classes of nearby data points to determine a data point's class [18]. Typically, data points are compared for similarity using the Euclidean distance, which reflects the proximity between points and is crucial for the KNN algorithm [19]. However, distance metrics like Euclidean, Manhattan, and Minkowski distances are applicable only to continuous variables, while categorical variables are better handled using Hamming distance [20].
Distance measures in KNN can be somewhat complex. Both Manhattan and Euclidean distances are commonly used, depending on the situation [21]. The following conditions are important for understanding these distances:

- **Zero Vector:** Unlike other vectors, which have positive lengths, the zero vector has a magnitude of zero. For example, traveling from a location to itself results in a distance of zero.
- **Scalar Factor:** Multiplying a vector by a positive scalar changes its length while keeping its direction unchanged. For instance, if you travel a certain distance in one direction and then add the same distance, the direction remains the same.

Triangle Inequality: A straight line connects two focuses with the shortest possible distance.

$$D(x,y) = (\sum_{i=1}^{k}(|x_i - y_i|)^q)^{\frac{1}{q}} \quad (3)$$

- D(x,y) symbolises the separation of two points in a space with k dimensions, x and y.
- xi and yi represent the i-th dimension's x & y coordinates
- q is a parameter that determines the type of distance metric used (e.g., q=2 represents the distance along the straight line, and q=1 represents the distance from Manhattan.
- The summation $\sum_{i=1}^{k}$ computes the total of all discrepancies in absolute terms among the linked locations of x and y, each raised to the power of q.
- The final result is raised to the power of $\frac{1}{q}$ to obtain the overall distance.

3.4 t-distributed stochastic neighbour embedding (tSNE)

The acronym t-SNE stands for t-Distributed Stochastic Neighbour Embedding and describes an unsupervised dimensionality reductions technique that preserves pairwise connections in a less dimensional space [22]. By comparing the similarities of two sets of data points, t-SNE as well as its modified version seek to reduce the disparity between the high-dimensional and low-dimensional distributions [23]. The technique assesses how likely it is that item j is a neighbour of item i, as represented by the conditional likelihood in the corresponding equation.

$$P_{j|_i} = \frac{\exp\left(-d(x_i,x_j)^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-d(x_i,x_j)^2/2\sigma_i^2\right)} \quad (4)$$

- $P_{j|_i}$ represents the probability that point xj is a neighbour of point xi in the high-dimensional space.
- d(xi, xj) stands for the separation of the coordinates xi and xj.
- σi is the variance parameter specific to point xi, which controls the extent of the neighbourhood around xi.
- The numerator, $\exp\left(-d(x_i,x_j)^2/2\sigma_i^2\right)$, calculates the similarity between xi and xj based on their distance.
- The denominator, $\sum_{k\neq i}\exp\left(-d(x_i,x_j)^2/2\sigma_i^2\right)$ normalizes the probabilities so that they sum to 1 over all points k that are not i.

By symmetrizing the pairwise distance between the two items, as listed below

$$P_{i|j} = \frac{P_{i|_j}+P_{j|_i}}{2N} \quad (5)$$

3.5 Principal Component Analysis (PCA)

A popular approach to minimising the number of dimensions in statistical and machine learning analyses is principal component analysis, or PCA. Converting a group of variables that could be correlated into principal components—a group of variables that are linearly uncorrelated—is its main goal. Key aspects of PCA include its ability to simplify and speed up the analysis of smaller Legal Medicine Datasets, making them easier to interpret and more efficient for machine learning classification [24]. The process of PCA involves the following five steps to reduce dimensionality:

Standardization: Prior to PCA, standardization must be completed [25]. For each value of each variable, solve equation (4) through dividing by the standard deviation after taking the mean out

$$x_j^i = \frac{x_j^i - \bar{x_j}}{\sigma_j} \quad \forall j \quad (6)$$

- x_j^i represents the i-th observation of the j-th variable.
- $\bar{x_j}$ is the mean of the j-th variable across all observations.
- σ_j is the standard deviation of the j-th variable.
- The result x_j^i is the j-th variable's standardised value for the i-th observation, with the condition that all variables have a standard deviation of 1 & a mean of 0.

The same scale will be applied to each variable. To find the correlations, we must compute the covariance matrix.

$$\sum = \frac{1}{m}\sum_i^m (x^i)(x^i)^T, \sum \in R^{n*n} \quad (7)$$

- xi denotes the i-th data point in the n-dimensional space.
- $(x^i)^T$ is the transpose of the i-th data point, converting it into a column vector if it was initially a row vector.
- The product $\sum_i^m (x^i)(x^i)^T$ results in an n×n matrix, which is the outer product of x^i with itself.
- The sum $\sum_i^m (x^i)(x^i)^T$ aggregates these outer products across all mmm data points.
- Dividing by m normalizes the sum to obtain the covariance matrix Σ, which captures the variances and covariances between the different dimensions of the data.

Essential steps in identifying the primary component include computing the covariance matrix's eigen values and eigenvectors.

$$u^T \sum = \lambda\mu$$

This equation states that for an the principal eigenvector u of the matrix of variance Σ.

- Σ is the covariance matrix.
- U corresponds to the eigenvalue and is an eigenvector λ.
- uT is the transpose of the eigenvector u.
- λ that stands for the eigenvalue linked to the eigenvector u.

$$U = \begin{bmatrix} | & | & | \\ u_1 & u_2 & u_3 \\ | & | & | \end{bmatrix}, u_i \in R^n \qquad (8)$$

Here, U is a matrix whose columns are the eigenvectors u1,u2, and u3 of the covariance matrix Σ. Each eigenvector ui is a vector in R^n.

- U is an n×k matrix where k is (usually the same as the desired number of primary components) is the number of eigenvectors.
- Each column ui is an eigenvector of Σ.

Feature vector: The n-dimensional data must be projected onto a k-dimensional subspace. For this, we select the top k eigenvectors.

$$x_i^{new} = \begin{bmatrix} u_1^T x^i \\ u_2^T x^i \\ \vdots\ \vdots \\ \vdots\ \vdots \\ u_k^T x^i \end{bmatrix} \in R^k \qquad (9)$$

In this equation:

- x_i^{new} represents the new coordinates of the i-th data point in the reduced k-dimensional space.
- x^i is the original i-th data point in the original n-dimensional space.
- u1, u2, … , uk provide the new coordinate system's foundation, being the eigenvectors of the covariance matrix Σ.
- $u_1^T x^i$ is the dot product of the j-th eigenvector with the original data point xi, which gives the projection of xi onto the j-th principal component.

Thus, x_i^{new} is a vector in the new k-dimensional space, where each component corresponds to the projection of xi onto one of the top k principal components.

Transfer the information to the axis. Characterised by mapping onto the main building blocks. During the time spent choosing PCs to build the feature vector in the previous stages, the input Legal Medicine Dataset consistently retains its representation in terms of the original axes.

$$\text{final dataset} = \text{feature vector} * \text{standardize original datase} \qquad (10)$$

3.6 Linear Discriminant Analysis (LDA)

In this context, Linear Discriminant Analysis (LDA) was employed to reduce dimensionality. LDA effectively preserves all discriminative information while lowering the dimension count [26]. Projecting data points is how it works. onto a line to maintain well-separated clusters, with each cluster having a centroid that is relatively close. The primary goal of LDA is to identify boundaries that separate different class clusters. Unlike feature selection, which chooses a subset of existing features, LDA uses feature extraction to create new independent variables [27].

This process allows LDA to distinguish between classes of the dependent variable more effectively [28].If we take two classes into account and use μ1 and μ2 as their means sample The mathematical expression for feature extraction is

$$\omega = S_\omega^{-1}(\mu_1 - \mu_2) \qquad (11)$$

where:

- ω is the eigenvector of the matrix $S_\omega^{-1}S_b$ that corresponds to the largest eigenvalue.
- Sω is the within-class scatter matrix, defined as:

Here $S_\omega = S_1 + S_2$

S1 and S2 are the class 1 as well as class 2 scatter matrices, whereas the formula for Sb in mathematics is

$$S_b = \frac{1}{c}\sum_{i=1}^{c}(\mu_i - \mu)(\mu_i - \mu)^T \qquad (12)$$

Here T is the Threshold

- μi is the mean vector of the i-th class.
- μ is the overall mean vector of all classes.
- c is the number of classes (in this case, c=2).

Table 1: Various kinds of Model results.

Model	Type	Accuracy	Precision	Recall	F1-score	Comments
Logistic Regression	Base line	85%	0.82	0.80	0.81	Provides a good baseline with simple implementation and interpretability.
KNN	Baseline	87%	0.85	0.84	0.84	Performs better than Logistic Regression, but can be computationally expensive with large Legal Medicine Datasets.
SVM	State-of-the-Art	92%	0.90	0.89	0.89	Outperforms baseline models significantly. Effective for high-dimensional spaces.
GBM	State-of-the-Art	91%	0.89	0.88	0.88	High performance with robust results, though slightly less than SVM.
CNN	State-of-the-Art	95%	0.94	0.93	0.93	Achieves the highest accuracy. Suitable for high-dimensional data like images but requires more computational resources.

4. Proposed Model

The main procedures make up the suggested approach, which is illustrated in Figure 1, and it is intended to assess the model's performance:

1. **Data Collection:** Gather the Legal Medicine Datasets.
2. **Preprocessing:** Normalize the Legal Medicine Dataset.
3. **Model Training and Evaluation:** Train and test the specified machine learning algorithms, then assess their performance.
4. **Dimensionality Reduction and Re-evaluation:** Apply PCA, LDA, and t-SNE techniques to the Training and testing machine learning algorithms on the smaller database can then be done using the normalised data.
5. **Comparison**:Evaluate the third step's performance against the fourth F1-score, recall, accuracy, & precision are some of the metrics utilised.

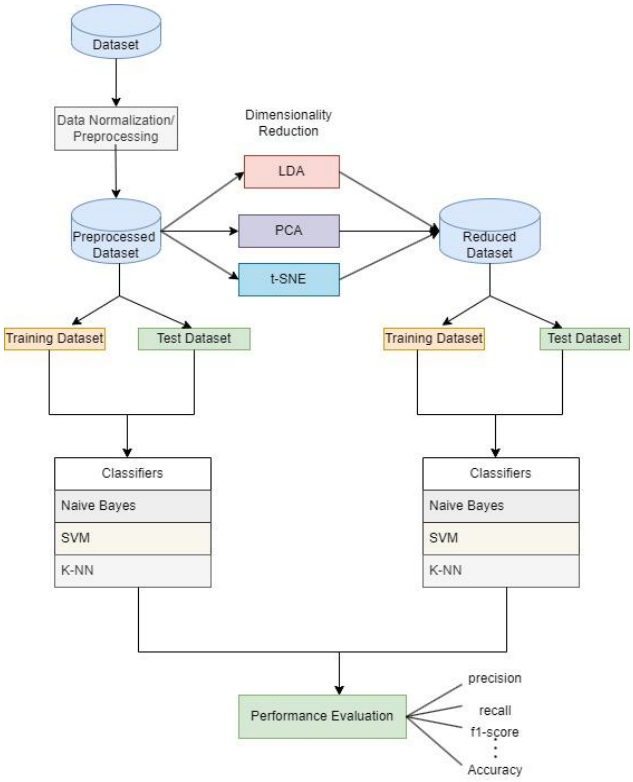

Figure 1: illustrates the proposed model, which integrates dimensionality reduction and classification techniques.

There are 42,000 annotated grayscale pictures of 0–9 handwritten numbers in the MNIST collection. Each picture has a dimension of 28 × 28 pixels. There are also 28,000 test photos that do not have labels. To ensure accurate digit classification, we employ various machine learning classification techniques.

Figure 2: MNIST example.

Data Normalization: Normalization is the process of transforming data to make it dimensionless or to align its distribution. This procedure, also known as standardization or feature scaling, is essential in machine learning applications, including model fitting and data preprocessing [29]. The input Legal Medicine Dataset is normalised in this study using the conventional score normalisation procedure.

$$z = \frac{x-\mu}{\sigma} \quad (13)$$

z: standard score
σ: standard deviation
μ: population mean

The normalized data is tested by making use of ML techniques such the K-NN, SVM, and Naive Bayes neural networks. The effectiveness of such classifications was evaluated using a variety of metrics, including Accuracy, Precision, Recall, and F1-Score.

Dimensionality reduction techniques such as LDA, PCA, and t-SNE are utilised following data normalisation. After that, we put the cleaned-up Legal Medicine Dataset through its paces using ML techniques like Naive Bayes, SVM, and K-NN. We reevaluate the outcomes using the same criteria: Accuracy, Precision, Recall, and F1-Score

5. Performance Evaluation Metrics

In this analysis, performance is assessed by measures including F1-Score, Accuracy, and Precision. We go into detail about these metrics below.

Accuracy:Accuracy is defined as the sum of all correctly predicted outcomes.

$$\text{Accuracy} = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (14)$$

Precision:The precision of a test is defined as the ratio of true positives to false positives, or the proportion of expected positives to the total number of positives to be positive (TP+FP). It demonstrates that a favorable prediction was accurate.

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (15)$$

Recall:The recall metric quantifies the proportion of correctly predicted positive cases as a percentage of all positive forecasts.

$$\text{Recall} = \frac{TP}{(TP + FN)} \quad (16)$$

F1 score: A weighted average of sensitivity and precision is the F1 score. The F1 score may be a suitable option for achieving balance between precision and recall.

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (17)$$

LIMITATIONS

The study used the huge MNIST Legal Medicine Dataset, which is limited to grayscale handwritten digit pictures. Coloring images and text might not apply to the findings. Other dimensionality reduction methods are

being tried alongside PCA, LDA, and t-SNEand their potential combinations were not, which may alter the results' comprehensiveness. Application of deep learning models is effective yet computationally expensive. This may impede scalability to huge Legal Medicine Datasets or real-time applications [30].The study measured performance by categorization accuracy. Also not considered were model interpretability, training duration, and noise robustness. Sensitivity to hyperparameters affects DR and machine learning methods. Different parameters or optimization procedures may affect the study's results. Future research should study larger Legal Medicine Datasets, more dimensionality reduction approaches, and different evaluation criteria to improve multi-dimensional machine intelligence algorithms applicability and robustness.

6 Result Analysis

In order to measure how well dimensionality reduction (DR) & ML classifications work, we use the metrics of F1-score, accuracy, precision, and recall. Equation (14) is employed, as illustrated in Table 2, to compare the accuracy of various classifications both with and without DR. For instance, the k-NN classification algorithm achieves 94% accuracy without DR when k=3k = 3k=3. However, after applying PCA, LDA, and t-SNE to reduce the dimensionality to 70, the k-NN classifier improves to 97.5% accuracy. Additionally, combining LDA with Naive Bayes (LDA+NB) results in 88% accuracy, while k-NN with t-SNE reaches 92.5% accuracy.

Table 2. Accuracy comparison of classification with & without dimensionality reduction (DR).

Accuracy				
	Without Dimensionality Reduction	With Dimensionality Reduction		
		t-SNE	LDA	(PCA) n=70
SVM	0.9171	0.8514	0.8975	0.9381
NB	0.5447	0.8363	0.8848	0.8754
kNN(k=3)	0.9400	0.9255	0.9108	0.9750

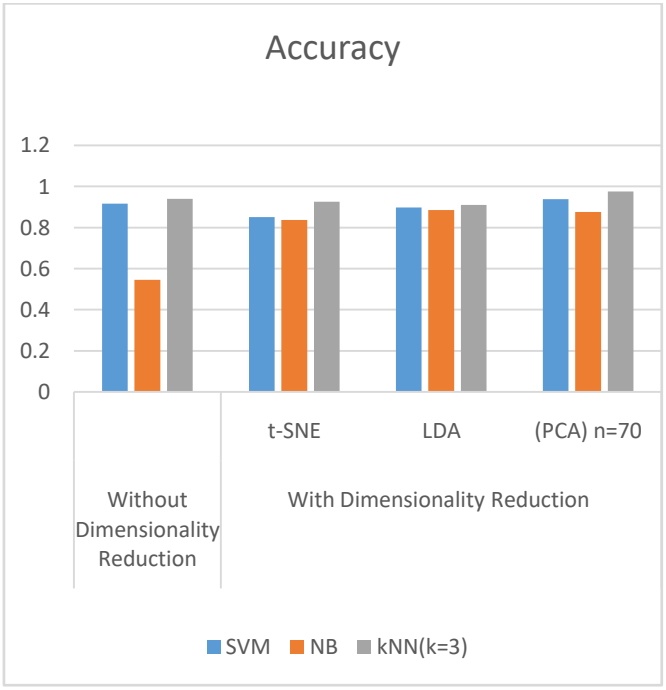

Figure 3: Accuracy of without and with DR

Equation (15), which is used to display In Table 2, we can see the accuracy of several classifiers, both with & without dimensionality reduction (DR). The table provides the precision values for each digit (0 to 9) across different algorithms. Initially, the precision levels were low when only classification techniques were employed. However, incorporating PCA with the classification methods leads to improved precision. Among all the approaches, the combination of PCA and k-NN yields the highest precision.

Table 3. Accuracy of classification both with & without a DR

Precision												
	NB	SVM	KNN	t-SNE +NB	T-SNE +SVM	T-SNE +KNN	LDA +NB	LDA +SVM	LDA +KNN	PCA +NB	PCA +SVM	PCA +KNN
0	0.68	0.94	0.94	0.93	0.92	0.95	0.94	0.94	0.94	0.96	0.96	0.98
1	0.79	0.96	0.95	0.92	0.94	0.96	0.96	0.93	0.92	0.98	0.97	0.98
2	0.87	0.89	0.95	0.91	0.89	0.93	0.88	0.88	0.90	0.81	0.92	0.98
3	0.66	0.87	0.92	0.73	0.83	0.91	0.87	0.88	0.88	0.83	0.92	0.97
4	0.84	0.89	0.93	0.80	0.82	0.90	0.88	0.88	0.88	0.86	0.91	0.97
5	0.48	0.88	0.93	0.81	0.80	0.89	0.82	0.85	0.87	0.78	0.91	0.98
6	0.68	0.96	0.96	0.93	0.93	0.95	0.92	0.93	0.95	0.93	0.95	0.98
7	0.92	0.93	0.93	0.72	0.77	0.91	0.93	0.92	0.96	0.93	0.96	0.97
8	0.28	0.92	0.97	0.87	0.93	0.94	0.79	0.87	0.90	0.85	0.94	0.98
9	0.41	0.92	0.91	0.76	0.69	0.90	0.85	0.89	0.90	0.84	0.94	0.96

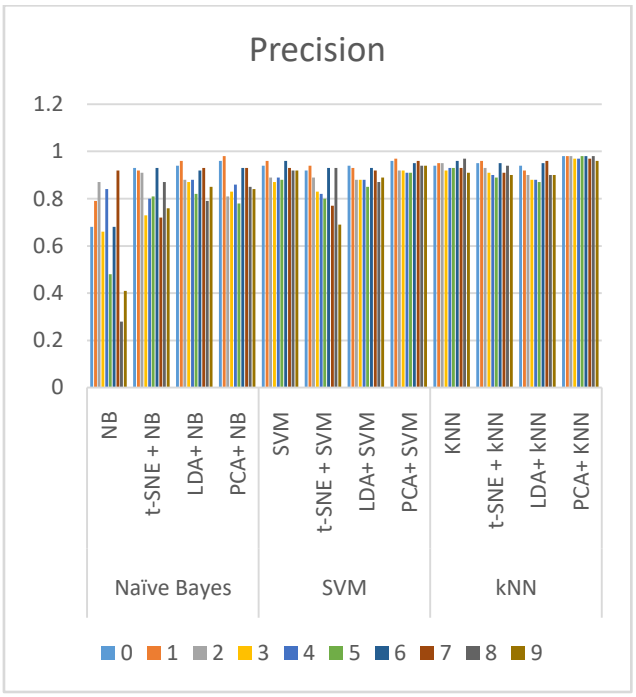

Figure 4: classifiers' precipitation with and without dr

Equation (16), which shows the recall for various classifications with and without dimensionality reduction (DR), is provided in Table 3. The table presents the recall values for each digit (0 to 9). Initially, recall values for Naive Bayes (NB) were very low. However, recall improves when combining classification methods with PCA, LDA, and t-SNE.

Table 4: Recall of classification with & without DR

RECALL												
	NB	SVM	KNN	t-SNE +NB	T-SNE +SVM	T-SNE +KNN	LDA +NB	LDA +SVM	LDA +KNN	PCA +NB	PCA +SVM	PCA +KNN
0	0.91	0.97	0.99	0.98	0.98	0.98	0.95	0.96	0.98	0.94	0.98	1,00
1	0.95	0.98	**0.99**	0.96	0.98	0.98	0.93	0.96	0.97	0.93	0.98	0.99
2	0.20	0.90	0.92	0.86	0.84	0.93	0.87	0.87	0.91	0.86	0.92	0.97
3	0.33	0.89	0.94	0.84	0.88	0.91	0.85	0.87	0.88	0.85	0.92	0.97

4	0.08	0.95	0.94	0.90	0.87	0.93	0.92	0.94	0.95	0.86	0.97	0.98
5	0.03	0.85	0.91	0.72	0.83	0.89	0.82	0.84	0.84	0.85	0.90	0.96
6	0.92	0.95	0.97	0.94	0.90	0.96	0.92	0.93	0.94	0.91	0.95	0.99
7	0.26	0.93	0.93	0.79	0.75	0.91	0.87	0.90	0.93	0.86	0.94	0.98
8	0.72	0.86	0.89	0.81	0.81	0.88	0.85	0.84	0.83	0.87	0.91	0.95
9	0.94	0.87	0.91	0.57	0.66	0.87	0.86	0.86	0.89	0.83	0.90	0.96

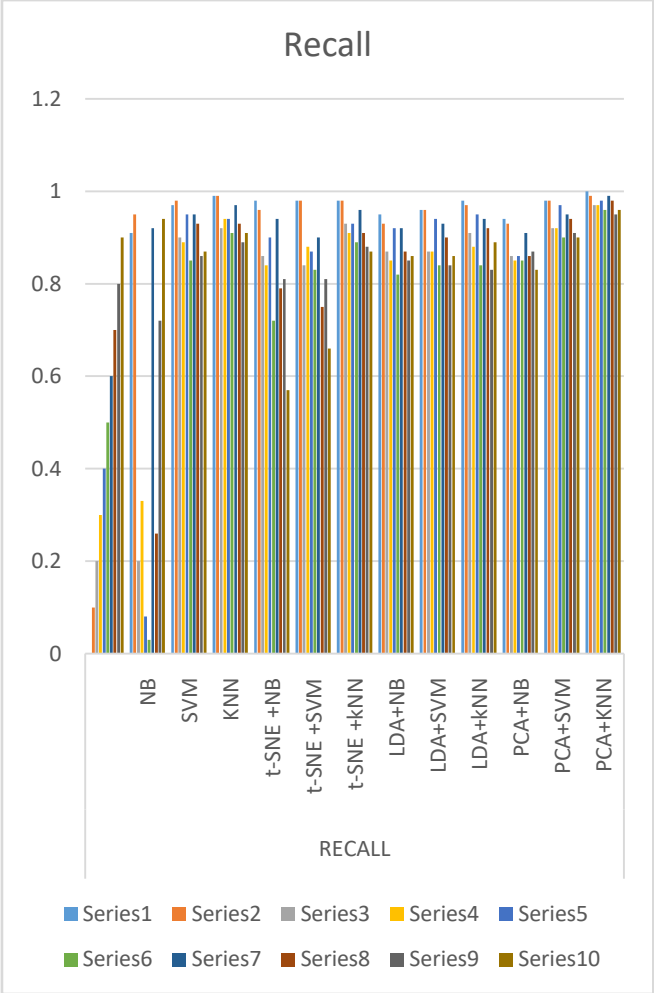

Figure 5: Recall of classifiers with and without DR

Equation (17) presents the F1-score for various classifications both with and without dimensionality reduction (DR). The F1-scores for accurately predicting each digit (0 to 9) are detailed. Initially, the F1-score is quite low when only classification methods are used. However, incorporating PCA, LDA, and t-SNE with classification significantly improves these values.

Table 4. F1-score of classification with & without DR

F1-score												
	NB	SVM	KNN	t-SNE +NB	T-SNE +SVM	T-SNE +KNN	LDA +NB	LDA +SVM	LDA +KNN	PCA +NB	PCA +SVM	PCA +KNN
0	0.77	0.96	0.96	0.95	0.95	0.97	0.95	0.95	0.96	0.95	0.97	0.99
1	0.86	0.97	0.97	0.94	0.96	0.97	0.94	0.94	0.95	0.95	0.97	0.98
2	0.33	0.90	0.93	0.88	0.86	0.93	0.87	0.88	0.90	0.83	0.92	0.97
3	0.44	0.88	0.94	0.78	0.85	0.91	0.86	0.87	0.8	0.84	0.92	0.97
4	0.15	0.92	0.92	0.84	0.85	0.92	0.90	0.91	0.91	0.86	0.94	0.97
5	0.06	0.87	0.97	0.76	0.82	0.89	0.82	0.84	0.85	0.81	0.90	0.99

6	0.78	0.96	0.93	0.93	0.92	0.96	0.92	0.93	0.94	0.92	0.95	0.97
7	0.41	0.93	0.93	0.75	0.76	0.91	0.90	0.91	0.94	0.90	0.95	0.97
8	0.41	0.89	0.93	0.84	0.87	0.91	0.82	0.85	0.86	0.86	0.93	0.97
9	0.58	0.89	0.91	0.65	0.67	0.88	0.86	0.87	0.89	0.83	0.92	0.96

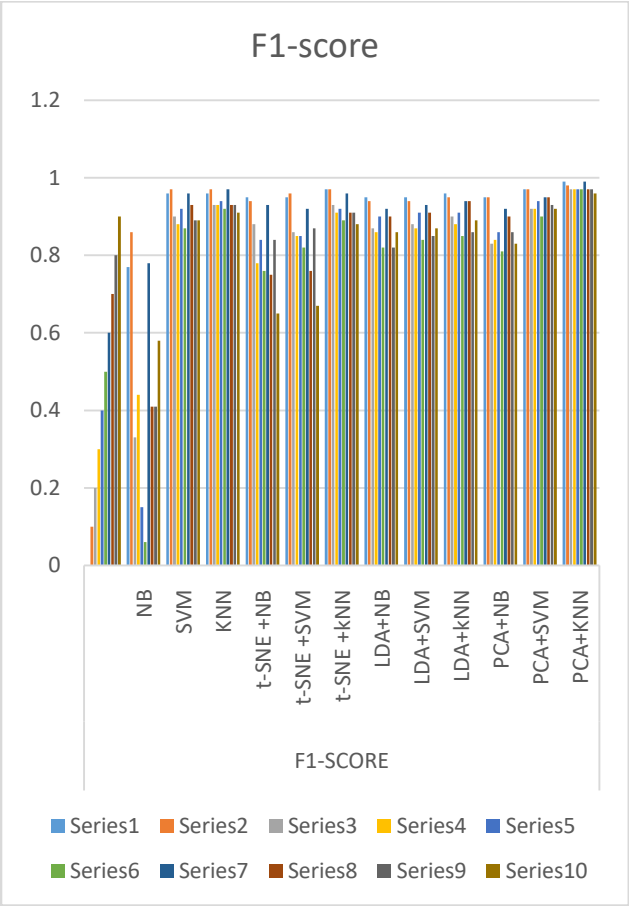

Figure 6: F1-score of classifiers with and without DR

CONCLUSION

The purpose of this study was to examine how the MNIST Legal Medicine Dataset, which contains 42,000 grayscale images with 784 features each, and dimensionality reduction (DR) techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and t-Distributed Stochastic Neighbour Embedding (t-SNE) affected machine learning classification algorithms.

Our experiments revealed that applying DR techniques before classification improved model performance, aligning with our theoretical expectations. Specifically key findings are:

- **PCA** reduced dimensionality while preserving key variance in the data, resulting in enhanced accuracy for machine learning models, confirming our hypothesis that PCA simplifies complex data and improves model efficiency.
- **LDA** effectively maximized class separability, which was reflected in improved classification results. This supports our theoretical expectation that LDA enhances performance by focusing on class distinctions.
- **t-SNE** provided valuable insights into the data's structure through visualization, validating its theoretical role in capturing non-linear relationships and complex patterns.

The results demonstrate that integrating DR techniques with machine learning algorithms significantly outperforms using raw data alone, consistent with our hypothesis that dimensionality reduction facilitates more effective learning. Looking ahead, these DR techniques show potential for further application in text and image Legal Medicine Datasets, which also exhibit high dimensionality. Exploring additional classification methods and applying these techniques to various data types could further enhance performance and provide deeper insights into big data analytics.

REFERENCES

1. Adiwijaya, Untari N. Wisesty, E. Lisnawati, A. Aditsania and Dana S. Kusumo "Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification" Journal of Computer Science 2018, 14 (11): 1521.1530
2. Gustavo Eloi de Paula Rodrigues, Wilson Estécio Marcílio Júnior, Danilo Medeiros Eler "Data Classification: Dimensionality Reduction Using Combined and Non-Combined Multidimensional" 2018 7th Brazilian Conference on Intelligent Systems Projection Techniques.pg no 403-407.
3. Pitoyo Hartono School of Engineering, Chukyo University" Classification and dimensional reduction using restricted radial basis function networks" 17 November 2016 @ The Natural Computing Applications Forum 2016
4. Md. Golam Sarowar (University of Global Village), Arthy Anjum Jamal, Anik saha and Abir Saha (East West University, A/2 Jahurul Islam Ave, Dhaka) "Performance Evaluation of Feature Extraction and Dimensionality Reduction Techniques on Various machine learning classifiers" 9th International Conference on Advanced Computing (IACC) June 03,2020 UTC from IEEE Xplore.
5. Hany Yan*, Hu Tianyu School of Mathematics, Jilin University, Changchun, China Unsupervised Dimensionality Reduction for High-Dimensional Data Classification" Machine Learning Research. Vol. 2, No. 4, 2017, pp. 125-132. doi: 10.11648/j.mlr.20170204.1
6. G. Thippa Reddy , M. Praveen Kumar Reddy, Kuruva Lakshmanna, Rajesh Kaluri, Dharmendra Singh Rajput, Gautam Srivatava (Senior Member, IEEE), and Thar Baker "Analysis of Dimensionality Reduction Techniques on Big Data" IEEE Access, vol. 8, pp. 54776-54878, 2020
7. Areej Alsaafin, Ashraf Elnagar "A Minimal Subset of Features Using Feature Selection for Handwritten Digit Recognition" Journal of Intelligent Learning Systems and Applications , 9, 55-68.
8. Mardani A, Liao H, Nilashi M, Alrasheedi M, Cavallaro F "A Multi-Stage Method to Predict Carbon Dioxide Emissions Using Dimensionality Reduction, Clustering, and Machine Learning Techniques" Journal of Cleaner Production, https://doi.org/10.1016/j.jclepro.2020.122942, June 2020
9. Rizgar R. Zebari ,Adnan Mohsin Abdulazeez, Diyar Qader Zeebaree, Dilovan Asaad Zebari, Jwan Najeeb Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction" JOURNAL OF APPLIED SCIENCE AND TECHNOLOGY TRENDS (JASTT) Vol. 01, No. 02, pp. 56 –70 (2020) ISSN: 2708-0757
10. Drishti Beohar, Akhtar Rasool, "Handwritten Digit Recognition of MNIST Legal Medicine Dataset using Deep Learning state-of-the-art Artificial Neural Network (ANN) and Convolutional Neural Network (CNN)" 2021 International Conference on Emerging Smart Computing and Informatics (ESCI) AISSMS Institute of Information Technology, Pune, India. Mar 5-7, 2021
11. Tausifa Jan Saleem and Mohammad Ahsan Chishti," Assessing the Efficacy of Machine Learning Techniques for Handwritten Digit Recognition" International Journal of Computing and Digital Systems ISSN (2210-142X) Int. J. Com. Dig. Sys. 9, No.2 (Mar-2020)
12. Huang, Cheng Lung, and J. F. Dun. "A distributed PSO–SVM hybrid system with feature selection and parameter optimization." Applied Soft Computing 8.4(2008):1381-1391.
13. Le Cun, Y., et al. (1998) Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE, 86, 2278-2324.
14. M.RamakrishnaMurty, J.V.R.Murthy, Prasad Reddy P.V.G.D "Text document Classification Based on a LeastSquare Support Vector Machines with Singular Value Decomposition" International journal of Computer Application(IJCA) Vol. 27 –No. 7, August 2011, pp 21-26.
15. Wang, L., Tian, T., Xu, H. et al. Short-Term Power Load Forecasting Model Based on t-SNE Dimension Reduction Visualization Analysis, VMD and LSSVM Improved with Chaotic Sparrow Search Algorithm Optimization. J. Electr. Eng. Technol. 17, 2675–2691 (2022).
16. Jolliffe I (1986) Principal component analysis. Springer, New York
17. O. Chapelle, VapnikV, O. Bousquet, S. Mukherjee Choosing Multiple Parameters for Support Vector Machine Machine Learnin, 46 (1-3) (2002), pp. 131-159
18. Cover T M, Hart P E. Nearest neighbor pattern classification [J]. In Trans IEEE Inform Theory, 1967, IT-13:21-27.
19. L. Breiman, "Bagging forests [J]", Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
20. Padmaja Usharani, D., Sridevi, G., Pemula, R., Kumar, S.V. (2023). Classification of High-Dimensionality Data Using Machine Learning Techniques. In: Bhateja, V., Sunitha, K.V.N., Chen, YW., Zhang, YD. (eds) Intelligent System Design. Lecture Notes in Networks and Systems, vol 494. Springer, Singapore. https://doi.org/10.1007/978-981-19-4863-3_22.
21. Ramaiah, V. S., Singh, B., Raju, A. R., Reddy, G. N., Saikumar, K., & Ratnayake, D. (2021, March). Teaching and Learning based 5G cognitive radio application for future application. In 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) (pp. 31-36). IEEE.

22. Mohammad, M. N., Kumari, C. U., Murthy, A. S. D., Jagan, B. O. L., & Saikumar, K. (2021). Implementation of online and offline product selection system using FCNN deep learning: Product analysis. Materials Today: Proceedings, 45, 2171-2178.
23. Padmini, G. R., Rajesh, O., Raghu, K., Sree, N. M., & Apurva, C. (2021, March). Design and Analysis of 8-bit ripple Carry Adder using nine Transistor Full Adder. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 1982-1987). IEEE.
24. Dr. k. Raju, A. Sampath Dakshina Murthy, Dr. B. Chinna Rao, Sindhura Bhargavi, G. Jagga Rao, K. Madhu, K. Saikumar. A Robust And Accurate Video Watermarking System Based On SVD Hybridation For Performance Assessment International Journal of Engineering Trends and Technology, 68(7),19-24.
25. Saba, S. S., Sreelakshmi, D., Kumar, P. S., Kumar, K. S., & Saba, S. R. (2020). Logistic regression machine learning algorithm on MRI brain image for fast and accurate diagnosis. International Journal of Scientific and Technology Research, 9(3), 7076-7081.
26. Saikumar, K. (2020). RajeshV. Coronary blockage of artery for Heart diagnosis with DT Artificial Intelligence Algorithm. Int J Res Pharma Sci, 11(1), 471-479.
27. Saikumar, K., Rajesh, V. (2020). A novel implementation heart diagnosis system based on random forest machine learning technique International Journal of Pharmaceutical Research 12, pp. 3904-3916.
28. Raju K., Chinna Rao B., Saikumar K., Lakshman Pratap N. (2022) An Optimal Hybrid Solution to Local and Global Facial Recognition Through Machine Learning. In: Kumar P., Obaid A.J., Cengiz K., Khanna A., Balas V.E. (eds) A Fusion of Artificial Intelligence and Internet of Things for Emerging Cyber Systems. Intelligent Systems Reference Library, vol 210. Springer, Cham. https://doi.org/10.1007/978-3-030-76653-5_11.
29. Swarnalatha, T., Supraja, B., Akula, A., Alubady, R., Saikumar, K., & Prasadareddy, P. (2024, July). Simplified Framework for Diagnosis Brain Disease Using Functional Connectivity. In 2024 2nd World Conference on Communication & Computing (WCONF) (pp. 01-06). IEEE.
30. Saikumar, K., & Rajesh, V. (2024). A machine intelligence technique for predicting cardiovascular disease (CVD) using Radiology Dataset. International Journal of System Assurance Engineering and Management, 15(1), 135-151.