P-ISSN: 2656-1743 | E-ISSN: 2656-1735 **JURNAL RISET INFORMATIKA**
DOI: https://doi.org/10.34288/jri.v7i3.375 Vol. 7, No. 3. Juni 2025
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

COMPARATIVE ANALYSIS OF DIMENSIONALITY REDUCTION FOR BREAST CANCER USING MACHINE LEARNING AND DEEP LEARNING

Fatimah Asmita Rani[-1], Duwi Lufita Marfiana[-2]

Computer Science / Information Technology
Nusa Mandiri University
fasmitarani@gmail.com[1], lufita.m98@gmail.com[2]

Abstract

Breast cancer is one of the leading causes of death among women worldwide. Accurate early detection is essential to improve patient survival rates. Therefore, an efficient and optimal detection method is needed. This study presents a comparative analysis between machine learning and deep learning models integrated with various dimensionality reduction techniques to improve the accuracy of breast cancer classification. The dimensionality reduction methods evaluated include Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), Independent Component Analysis (ICA), and Linear Discriminant Analysis (LDA). This study uses a dataset from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), which includes genetic and clinical data of breast cancer patients. Several classification algorithms are used in the evaluation, including Logistic Regression, Support Vector Machines (SVM), and Convolutional Neural Networks (CNN). Model performance is analyzed based on accuracy, precision, recall, and F1-score metrics. The results show that the LDA technique consistently produces better classification performance compared to other dimensionality reduction methods on various Machine Learning and Deep Learning models. The importance of choosing the right dimensionality reduction method in increasing the effectiveness of learning algorithms and more optimal, especially in the context of complex and high-dimensional medical data. The implications of this study can be used to develop a smarter decision support system in breast cancer diagnosis.

Keywords: *Breast Cancer, High Dimensionality, Machine Learning, Deep Learning*

Abstrak

Kanker payudara merupakan salah satu penyebab kematian utama pada wanita di seluruh dunia. Deteksi dini yang akurat sangat penting untuk meningkatkan angka kesintasan pasien. Oleh karena itu, diperlukan metode deteksi yang efisien dan optimal. Penelitian ini menyajikan analisis perbandingan antara model machine learning dan deep learning yang diintegrasikan dengan berbagai teknik reduksi dimensionalitas untuk meningkatkan akurasi klasifikasi kanker payudara. Metode reduksi dimensionalitas yang dievaluasi meliputi Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), Independent Component Analysis (ICA), dan Linear Discriminant Analysis (LDA). Penelitian ini menggunakan dataset dari Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), yang meliputi data genetik dan klinis pasien kanker payudara. Beberapa algoritma klasifikasi digunakan dalam evaluasi, meliputi Logistic Regression, Support Vector Machines (SVM), dan Convolutional Neural Networks (CNN). Kinerja model dianalisis berdasarkan metrik akurasi, presisi, recall, dan F1-score. Hasil penelitian menunjukkan bahwa teknik LDA secara konsisten menghasilkan kinerja klasifikasi yang lebih baik dibandingkan dengan metode reduksi dimensionalitas lainnya pada berbagai model Machine Learning dan Deep Learning. Pentingnya pemilihan metode reduksi dimensionalitas yang tepat dalam meningkatkan efektivitas algoritma pembelajaran dan lebih optimal, terutama dalam konteks data medis yang kompleks dan berdimensi tinggi. Implikasi dari penelitian ini dapat digunakan untuk mengembangkan sistem pendukung keputusan yang lebih cerdas dalam diagnosis kanker payudara.

Kata kunci: Kanker Payudara, Dimensi Tinggi, Pembelajaran Mesin, Pembelajaran Mendalam

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v7i3.375
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

INTRODUCTION

Breast cancer is the second leading cause of death among women worldwide, highlighting its profound impact on public health (Islam et al., 2020). According to data reported in *Cancer Statistics, 2018* (Chugh et al., 2021), approximately 2.25 million people worldwide are living with cancer. Each year, more than 1,157,294 new cancer cases are registered, and nearly 784,821 cancer-related deaths are reported. The risk of dying from cancer is estimated to be 7.34% for males and 6.28% for females. Among males, approximately 25% of cancer-related deaths are attributed to oral cavity cancer and lung cancer. Meanwhile, for females, breast cancer and oral cavity cancer together account for 25% of all reported cancer cases (Chugh et al., 2021).

Figure 1 provides a detailed breakdown of cancer statistics for the year 2018. The data reveal that breast cancer is the most prevalent type of cancer among women, representing about 14% of all cancer cases in this demographic. According to the *Globocan 2018* study (Chugh et al., 2021), it is estimated that there were approximately 162,468 new cases of breast cancer and 87,090 deaths caused by this disease in the same year. These statistics highlight the substantial global burden of cancer, particularly for the most common types, such as breast cancer in women and lung cancer in men.

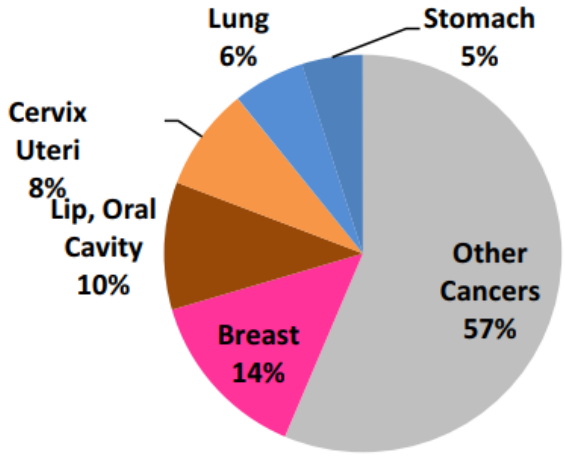

Figure 1. Cancer Statistics 2018(Chugh et al., 2021)

In 2019, an estimated 268,600 new cases of invasive breast cancer were expected to be diagnosed in women in the United States. Additionally, approximately 62,930 new cases of non-invasive breast cancer were projected, reflecting the widespread prevalence of this disease. These figures emphasize the critical importance of early detection, effective treatment, and ongoing research to reduce breast cancer incidence and mortality rates (S. A. Mohammed et al., 2020). Early detection and timely diagnosis are the only effective ways to prevent mortality caused by this disease. According to estimates from the European Cancer Information System (ECIS), the mortality rate of breast cancer in European countries was 13.4%. Furthermore, in 2020, the national costs associated with breast cancer were significantly higher compared to other diseases. These findings highlight the critical need for prevention efforts, increased awareness, and improved access to healthcare services to reduce the social and economic burden of breast cancer (Dafni et al., 2019).

The American Cancer Society published a report indicating that the breast cancer death rate for all age groups is 41,760 for females and 500 for males. Meanwhile, the Globocan project report reveals that in India, 162,468 new cases of breast cancer were detected, with a death rate of 87,090 per year. Medical image analysis research groups are continually focused on developing new screening technologies and improving early cancer detection. Automated and early detection of cancer becomes not only cost-effective but also less time-consuming compared to traditional methods. This offers significant potential to improve access to healthcare services, particularly in resource-limited countries, and to reduce the overall societal impact of breast cancer (Muduli et al., 2022).

Research related to breast cancer studied by Jiande Wu and Chindo Hicks (Wu & Hicks, 2021) investigates the application of machine learning (ML) algorithms to classify breast cancer into Triple Negative Breast Cancer (TNBC) and non-TNBC using RNA-Sequence data from The Cancer Genome Atlas (TCGA). The study involved preprocessing the data through quality control, normalization (using LIMMA and edgeR in R), and feature selection, which identified 5,502 significant differentially expressed genes (DEGs). Four ML algorithms Support Vector Machines (SVM), K-Nearest Neighbor (kNN), Naïve Bayes (NGB), and Decision Tree (DT) were evaluated using various subsets of genes, with their performance assessed based on accuracy, sensitivity, specificity, precision, and F1 score. Among these, SVM achieved the best performance, with an accuracy of 90%, a recall of 87%, and a specificity of 90%, particularly when using the top 256 genes. The study's strengths include its novelty as the first to apply ML models specifically to classify TNBC and non-TNBC using RNA-seq data and its clinical relevance in identifying high-risk patients for early intervention

and personalized treatment. However, the research faced limitations, including imbalanced data, with TNBC samples significantly fewer than non-TNBC, which may affect generalizability. Additionally, it did not address the classification of subtypes within TNBC and non-TNBC, and only four ML algorithms were evaluated, leaving out more advanced approaches such as ensemble methods or neural networks (Wu & Hicks, 2021).

Noreen Fatima et al (Fatima et al., 2020) provides an extensive review of various machine learning (ML) and deep learning (DL) algorithms applied to breast cancer prediction. The authors evaluated techniques such as Support Vector Machines (SVM), Decision Trees (DT), K-Nearest Neighbors (KNN), Naïve Bayes (NB), Random Forests (RF), Convolutional Neural Networks (CNN), Autoencoders, and Recurrent Neural Networks (RNN), along with ensemble methods like Bagging, Boosting, and Stacking. The study analyzed these algorithms based on metrics like accuracy, sensitivity, and specificity using datasets such as the Wisconsin Diagnostic Breast Cancer (WDBC) and The Cancer Genome Atlas (TCGA). SVM consistently demonstrated superior performance, achieving accuracies up to 99.68% on specific datasets, while ensemble methods effectively handled imbalanced data, and CNNs excelled in image-based cancer detection with accuracies up to 98.9%. The paper's strengths include its comprehensive analysis of a broad range of algorithms, practical relevance using real-world datasets, and a comparative approach that highlights the strengths and limitations of each method. However, the study's limitations lie in its heavy reliance on specific datasets, which may reduce the generalizability of its findings, and its focus on existing studies without proposing novel methods or insights. Additionally, the paper lacks a detailed exploration of hybrid models or advanced DL architectures. The authors conclude that SVM and ensemble techniques are the most reliable for breast cancer prediction, while CNNs are ideal for image-based tasks (Fatima et al., 2020).

Mohammed Amine Naji et al (Naji et al., 2021) explores the application of machine learning (ML) algorithms to predict and diagnose breast cancer using the Wisconsin Breast Cancer Diagnostic Dataset (WBCD). The dataset comprises 569 instances, with 62.74% categorized as benign and 37.26% as malignant. Five ML algorithms Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree (C4.5), and K-Nearest Neighbors (KNN)—were evaluated. The methodology involved data preprocessing (cleaning, attribute selection, feature extraction) and splitting the dataset into training (75%) and testing (25%) sets. Models were assessed using metrics such as accuracy, precision, sensitivity, F1 score, and AUC, with experiments conducted in the Python-based Anaconda environment utilizing the Scikit-learn library. Among the models, SVM achieved the highest accuracy (97.2%) and demonstrated superior performance across all metrics, including a precision of 0.98, sensitivity of 0.94, and an AUC of 0.966. The study's strengths include a comprehensive comparison of algorithms, SVM's demonstrated reliability, and a clear methodology. However, it is limited by its reliance on a single dataset, lack of advanced models like deep learning, and insufficient handling of class imbalance in the dataset (Naji et al., 2021).

Based on the reviewed studies, it is evident that machine learning and deep learning techniques (Fatima et al., 2020) hold great potential for breast cancer classification, particularly when enhanced with advanced preprocessing methods and robust algorithms. For future research, employing Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees, Random Forests, Multi-Layer Perceptrons (MLP), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks. These methods, when combined with high-dimensionality data, StandardScaler for feature normalization, and Synthetic Minority Over-sampling Technique (SMOTE) to address data imbalance, can further improve classification accuracy and generalizability.

The use of high-dimensional data allows for capturing complex patterns, while SMOTE effectively handles class imbalance, which is often a challenge in medical datasets. Incorporating deep learning architectures such as CNNs and LSTMs provides an opportunity to leverage their ability to learn intricate spatial and sequential data features, respectively.

RESEARCH METHODS

In this research, author used preprocessing techniques, and machine learning models used to develop a robust breast cancer classification framework. The proposed methodology includes data normalization using StandardScaler to standardize feature scales and the application of Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance, a common issue in medical datasets.

For the classification task, a comprehensive set of machine learning and deep learning models were selected: Logistic Regression, Support Vector

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v7i3.375
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees, Random Forests, Multi-Layer Perceptrons (MLP), Convolutional Neural Networks (CNN), and Long Short-Term Memory networks (LSTM). These algorithms were chosen based on their proven effectiveness in previous breast cancer studies and their capability to handle complex and diverse data types.

The approach to be used in this study is experimental, with the research flowchart as shown in Figure 2.

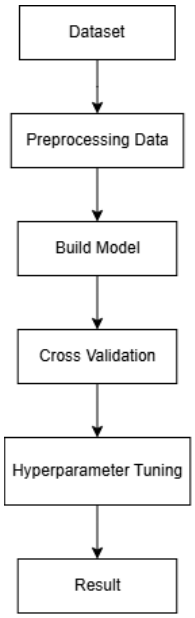

Figure 2. Research Flow Diagram

The stages in Figure 2 represent the research procedure, which includes the following steps that can be explained as follows:

1. Dataset

The dataset used in this study was obtained from The Breast Cancer Gene Expression Profiles (METABRIC) dataset provides comprehensive clinical attributes, mRNA z-score levels, and gene mutation data for a total of 1,904 patients (Pereira et al., 2016). This dataset is a product of the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), a collaborative research initiative between Canada and the United Kingdom. The dataset includes targeted sequencing data for 1,980 primary breast cancer samples and integrates both clinical and genomic information. This extensive data collection effort was led by Professor Carlos Caldas from the Cambridge Research Institute and Professor Sam Aparicio from the British Columbia Cancer Centre in Canada. The dataset is publicly accessible through cBioPortal and was first published in Nature Communications in 2016, as detailed in the study by Pereira et al (Pereira et al., 2016).
Below are clinical attributes in the dataset:

Table 1. Clinical Attributes Dataset

Name	Type
patient_id	object
age_at_diagnosis	float
type_of_breast_surgery	object
cancer_type	object
cancer_type_detailed	object
cellularity	object
chemotherapy	int
pam50_+_claudin-low_subtype	object
cohort	float
er_status_measured_by_ihc	float
er_status	object
neoplasm_histologic_grade	int
her2_status_measured_by_snp6	object
her2_status	object
tumor_other_histologic_subtype	object
hormone_therapy	int
inferred_menopausal_state	object

The genetic component of the dataset includes two key types of data: mRNA z-score levels and gene mutation data. Specifically, it provides mRNA z-score levels for a total of 331 genes and mutation information for 175 genes, offering a detailed view of the genetic landscape associated with breast cancer.

2. Preprocessing Data

The data preprocessing begins with identifying categorical columns using the select_dtypes method, which filters columns with the "object" data type. Certain irrelevant columns, such as patient_id (a unique identifier) and death_from_cancer (an auxiliary target variable), are excluded from this list. The patient_id column is removed as it does not contribute to prediction, and rows with missing values are dropped to ensure the data is clean.

Next, the dataset is split into features (X) and the target variable (y). The features include all columns except the primary target, overall_survival, and other irrelevant columns like death_from_cancer. After this separation, the features are normalized using StandardScaler to ensure that the data has a mean of 0 and a standard deviation of 1. Following normalization, dimensionality reduction is performed using

Independent Component Analysis (ICA), Principal Component Analysis (PCA), t-SNE and Linear Discriminant Analysis (LDA) to simplify the data without losing important information. By setting the number of components to 50, the data is transformed into a lower-dimensional representation. This step helps reduce computational complexity while preserving the core variance in the dataset.

The final data preprocessing is step addresses class imbalance in the target variable. The Synthetic Minority Oversampling Technique (SMOTE) is employed to generate synthetic data for the minority class, ensuring a balanced class distribution.

Below is explain of preprocessing method:

A. Independent Component Analysis (ICA)

Independent Component Analysis (ICA) was originally introduced to address blind signal separation problems (Fleuret et al., 2021). Its primary objective is to project data into a lower-dimensional space while maximizing independence among the components. From a theoretical standpoint, suppose we have m observations $(x_1, \dots, x_m)$, where each observation is a mixture of n independent components. The relationship can be expressed as:

$$x_i = \sum_{j=1}^{n} a_{i,j}\, s_j$$

Here, $a_{i,j}$ represents the mixing coefficients, and s_j denotes the independent components. This can be reformulated in matrix notation as:

$$x = As$$

where x and s are random vectors for observations and independent components, respectively, and A is the mixing matrix (Fleuret et al., 2021).

B. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is one of the most widely used dimensionality reduction techniques in data mining. It aims to identify data points with the highest variance by leveraging statistical methods. Through PCA, redundant and irrelevant features are eliminated, while the remaining features are reorganized into a new coordinate system called the principal space. This process enhances the visibility and interpretability of the data (Mahmoudi et al., 2021).

C. t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear, unsupervised, manifold-based feature extraction method designed to map high-dimensional data to lower dimensions (typically 2 or 3), while preserving the essential structure of the original dataset. t-SNE is primarily used for data exploration and visualization, offering insights into how high-dimensional data is organized. Unlike other feature extraction approaches, which often struggle to visualize high-dimensional data effectively, t-SNE excels at preserving both local and global structures of the data (Anowar et al., 2021).

The t-SNE process begins by applying Stochastic Neighbor Embedding (SNE) to the dataset. This step transforms high-dimensional Euclidean distances into conditional probabilities, representing the similarity between every pair of data points. For two points x_a dan x_b their similarity is expressed by the conditional probability $p_{a|b}$ calculated as:

$$p_{a|b} = \frac{\exp\left(-\frac{\left\|x_b - x_a\right\|^2}{2\sigma^2}\right)}{\sum_{k \neq a} exp\left(-\frac{\left\|x_k - x_a\right\|^2}{2\sigma^2}\right)}$$

Here, σ^2 represents the variance of the Gaussian distribution centered around x_b, which varies based on data density.

D. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is typically regarded as a linear and supervised feature extraction method, although some researchers classify it as a linear classifier. The primary objective of LDA is to identify a new feature space where data can be projected to maximize the separability between classes. From the d independent features in a dataset, LDA derives k new independent features that best distinguish the dependent classes. The number of components produced by LDA is always less than or equal to the number of classes minus one $(k \leq classes - 1)$ (Anowar et al., 2021).

LDA starts by constructing two scatter matrices:

1. The between-class scatter matrix $\left(S_{Mb}\right)$, which measures the variance between the means of different classes.
2. The within-class scatter matrix (S_{Mw}), which calculates the variance within each

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v7i3.375
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

class by measuring the distance between individual data points and their respective class means.

These matrices are computed as follows:

$$S_{Mb} = \sum_{k=1}^{m} N_k (\mu_k - \mu)(\mu_k - \mu)^T$$

$$S_{Mw} = \sum_{k=1}^{m} \sum_{x=1}^{n} (x - \acute{\mu}_k)(x - \acute{\mu}_k)^T$$

Here, m represents the number of classes, μ is the overall mean, μ_k and N_k are the mean and size of class k, respectively, and $\acute{\mu}_k$ is the class mean vector.

While LDA is a powerful tool for reducing dimensionality and improving class separability, it has a notable limitation in binary classification problems: it generates only one new feature, regardless of the number of original features in the dataset (Anowar et al., 2021).

E. Synthetic Minority Oversampling Technique (SMOTE)

As the minority class is over-sampled by increasing amounts, the etc. is to identify similar but more specific regions in the feature space as the decision region for the minority class. The synthetic minority over-sampling technique (SMOTE) is utilized to classify imbalanced datasets. This technique synthesizes new samples of the minority class to balance a dataset by re-sampling the instances of the minority class (Brandt & Lanzén, 2020).

Synthetic Minority Over-sampling Technique (SMOTE) is an oversampling method used to increase the number of samples that are contained in the minority class. The operational principle of SMOTE has its fundaments in the generation of synthetic data by using the K-nearest neighbors method. By using such an approach, it is possible to increase the number of members of the minority class by using the data generated in the feature space (Glucina,Matko; Lorencin,Ariana; Andelic,Nikola; Lorencin, 2023). SMOTE solves the problem of data imbalance by artificial linear interpolation of minority class samples so that it can prevent overfitting caused by random over-sampling (Wang et al., 2023)

3. Model

Classification is one of the most common tasks of machine learning and is a problem of classification unknown instance in one of the pre-offered categories classes. The important observation in classification is that target functions are discrete. In machine learning and statistics, classification is defined as training a system with labeled dataset to identify a new unseen dataset to which class it belongs. Recently, there is enormous growth in data and, unfortunately, there is lack of quality labeled data (R. Mohammed et al., 2020).

For this purpose, we have organized the paper in the following way. In the second part of this paper, we present evaluation of classification models, in the third part of the paper we present measures for the evaluation of classification models. In the last part of the paper, we discuss the results and give directions for further research.

A. Logistic Regression

Logistic Regression (LR), as a linear model, tends to overfit in scenarios where the dimensionality of the feature space is greater than the number of data points. This is because the model attempts to fit the noise in the data rather than capturing the underlying patterns (Zaidi & Al Luhayb, 2023).

In this research, Logistic Regression is applied as a classification model that predicts outcomes based on specific independent variables or predictors. The dependent variable in Logistic Regression is typically binary, representing categories such as "success" or "failure." In the context of this study, it is used to classify stroke patients into outcome groups, such as "good recovery" or "poor recovery," based on relevant clinical and demographic predictors. This approach highlights the model's utility for binary classification problems, while acknowledging and addressing its limitations through robust evaluation methods (Bielewicz et al., 2020).

For this model, the probability of a stroke outcome, given as, in relation to *n* prognosis-related covariates, X_1, X_2, X_3, is given as follows:

$$logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right)\beta_0 + \beta_1 X_1 + \cdots + \beta_i X_i$$
$$= \beta_0 + \sum_{i=1}^{n} \beta_i X_i$$

Here, β_i is the intercept; β_1, β_2, and β_i are the regression coefficients for the predictive variables X_1, X_2 and X_i. The regression coefficients indicate the contributions of the predictive variables to the results. The regression coefficients and their standard deviations are generally based on the least-squares fitting method (Qu et al., 2022).

B. SVM

Another classification technique utilized in this research is Support Vector Machines (SVM). The operation of SVM revolves around identifying a

boundary, or hyperplane, that separates data points into distinct classes. This boundary is determined using a kernel function, which transforms the original data into a higher-dimensional space to make it more separable. In this research, multiple kernel functions were explored, including linear, polynomial, and radial basis function (RBF) kernels, along with varying kernel-specific parameters such as coefficients and scaling values, to identify the most effective configuration for classification (Mohammadi et al., 2021).

SVM is a supervised machine learning method designed for classification tasks. It works by constructing a hyperplane that distinctly separates data points belonging to different classes, such as positive versus negative opinions. The algorithm focuses on maximizing the margin, which is the distance between the hyperplane and the closest data points from each class, known as support vectors. The rationale behind maximizing the margin is to enhance the model's generalization ability, reducing its likelihood of misclassification when applied to new, unseen data (Mohammadi et al., 2021).

Moreover, the hyperparameter tuning process in SVM is crucial for performance. Parameters such as the penalty term C control the trade-off between achieving a low training error and a large margin. A high C value prioritizes classifying training examples correctly, potentially leading to overfitting, while a low C value emphasizes a larger margin, potentially underfitting the data. Similarly, the choice of kernel parameters, like the degree for polynomial kernels or the gamma parameter for RBF kernels, significantly affects the classifier's ability to capture complex patterns in the data.

In summary, SVM is a powerful classification method, particularly suited for high-dimensional datasets. Its focus on maximizing the margin and ability to handle non-linear data through kernel functions makes it a robust choice for various classification problems. In this study, the evaluation of multiple kernel functions and parameter combinations allowed the identification of optimal settings, enhancing classification accuracy and minimizing generalization errors (Mohammadi et al., 2021).

C. KNN

The K-Nearest Neighbor (K-NN) algorithm is a classification technique that operates by referencing previously labeled data to classify new data points (Uddin et al., 2022). The algorithm functions by calculating the shortest distance between the test sample and each training sample to identify the k k-nearest neighbors. Once these neighbors are determined, the majority class among them is selected as the predicted class for the test instance. The proximity of neighbors is commonly assessed using the Euclidean distance, although other distance metrics can also be applied depending on the data's characteristics.

KNN algorithm is using Euclidean distance calculations. The formulation is as follows:

$$euc = \sqrt{(\sum_{i=1}^{n}(pi - qi)^2)}$$

pi = sample data / training data
qi = test data / testing data
i = data variable
n = data dimension

D. Decision Tree

A tree is a data structure composed of vertices (nodes) and edges (connections between nodes). Nodes in a tree are categorized into three types: the root node, which serves as the starting point of the tree; internal or branch nodes, which connect to other nodes; and leaf nodes, which represent the endpoints of the tree and have no further branches. In this structure, the root and internal nodes are labeled with attribute names, the edges are marked with possible attribute values, and the leaf nodes are identified by distinct classes or outcomes (Luo et al., 2023). Decision trees are a data mining technology that has been widely applied in the healthcare field.

$$H(S) = -\sum_{i=1}^{C} P_I log_2(p_i)$$

where c_i is the number of classes, and 〖p〗_i s the proportion of examples in class i.

$$IG(S, A) = H(S) - \sum_{v \in values}^{C} \frac{|s_v|}{|s|} H(S_v)$$

Where A is an attribute, S is the current set of examples, Sv is the subset of S for which attribute A has value v.

$$Gini(S) = 1 - \sum_{i=1}^{C} P_i^2$$

Where c is the number of classes, and pi, is the proportion of examples in class i

$$MSE(S) = \frac{1}{|S|} \sum_{i \in S} (y_i - \overline{y)^2}$$

Where S is the set of examples, yi is the target value example I, and y is the mean of the target values in S (Luo et al., 2023).

E. Random Forest

Breiman introduced the Random Forest (RF) algorithm as an ensemble learning technique to address both regression and classification problems. Ensemble learning leverages multiple

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v7i3.375
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

models to solve a single problem, enhancing overall accuracy. By aggregating multiple models, ensemble methods reduce variance, particularly with unstable classifiers, leading to improved outcomes. Random Forest operates using two key ensemble methods: Bagging and Boosting. Bagging (Bootstrap Aggregating) creates diverse training subsets by sampling the original training data with replacement. Boosting, on the other hand, sequentially combines weak learners into a robust model by correcting errors from prior models to strengthen overall performance (Sheykhmousa et al., 2020).

Random Forest is essentially a combination of multiple decision trees. It enhances the classification accuracy of an individual tree classifier by integrating the bagging method with a randomized selection of data partitions at each decision tree node. This approach mitigates overfitting and improves generalization. A decision tree with M leaves splits the feature space into M regions Rm, 1 ≤ m ≤ M. For each tree, the prediction function f(x) is defined as Formulas (1) and (2):

$$f(x) = \sum_{m=1}^{M} c_m \prod (x, R_m)$$

where M is the number of regions in the feature space, Rm is a region corresponding to m, cm is a constant corresponding to m (Dewi & Chen, 2019).

F. MLP

The Multilayer Perceptron (MLP) is an artificial neural network (ANN) classification algorithm that requires normalized data as input. MLPs are a powerful class of nonlinear statistical models composed of multiple layers of nodes in a directed graph, with each layer being fully connected to the next. There are three types of layers: input, hidden, and output layers. All nodes, except for those in the input layer, are neurons (processing elements) with nonlinear activation functions (Orrù et al., 2020).

Given input data x_i (for i=1,2,…,N), the output y of the neural model is calculated as follows:

$$y = f(W^T x + b)$$

Where:

- f is the activation function
- N is the number of neurons
- W are the weights of the ANN model
- b is the bias vector.

For binary classification, the output of an MLP is a value between 0 and 1, which can be interpreted as the probability of the positive target class. A parameter optimization loop was employed to tune the model's hyperparameters to maximize precision and recall for class "1," (Orrù et al., 2020).

G. CNN

Convolutional Neural Networks (CNN) have emerged as a powerful deep learning technique for solving complex tasks across various domains in recent years. As a result, their use has significantly increased in multiple areas of computer science and engineering. The architecture of a CNN model typically includes various components such as convolutional layers (CL), activation functions (AF), max-pooling, fully connected layers (FCL), dropout layers, and a softmax function for classification tasks (Watanobe et al., 2023).

The convolutional layers learn the features of the code from the input sequences, and their outputs are passed through activation functions like ReLU or LeakyReLU. The ReLU and LeakyReLU activation functions are defined as:

$$ReLU\left(f(z)\right) = \max(0, z)$$

$$LeakyReLU\left(f(z)\right) = \max(\alpha z, z)$$

Where z represents the input, and α is a small magnitude constant.

H. LSTM

LSTM (Long Short-Term Memory) is a type of artificial neural network designed to handle sequential data, such as time series, audio, or text. It is particularly effective for processing data with long-term dependencies, where the output at a given time step depends on information from previous time steps. LSTM networks can retain information over longer periods using memory cells, input gates, output gates, and forget gates. These gates regulate the flow of data into and out of the memory cells, allowing the network to selectively store and retrieve information as needed. LSTMs are widely used for tasks such as language translation, speech recognition, and stock price prediction (Gülmez, 2023).

4. Cross Validation

K-fold cross-validation is a commonly recommended method for addressing the challenge of obtaining an appropriate classifier and reliable performance estimates, particularly when working with small datasets. In K-fold cross-validation, the entire dataset D is randomly divided into K equally sized subsets (or folds) $D_1, D_2 \ldots . D_k$, such that the union of all these folds equals D and their pairwise intersections are empty (Oyedele, 2023).

$$Performance\ Measure = \frac{1}{K}\sum_{i=1}^{k} Evaluate\ (Model, Fold)$$

k is the number of folds. Evaluate Model (Model, Fold) Evaluate Model (Model, Fold) is the performance metric (accuracy, precision, recall, etc.) of the model when evaluated on the test set (Fold) during the i iteration (Oyedele, 2023).

5. Hyperparameter Tuning

Hyperparameters are key values that influence the behaviour of machine learning algorithms and directly affect how the model learns. These values, unlike model parameters that are learned during training, are predefined before the training process begins. Examples of hyperparameters include learning rate, number of hidden layers in neural networks, and the number of trees in a random forest model. Optimizing these hyperparameters is crucial for improving model performance.

One common technique for hyperparameter optimization is Grid Search, which exhaustively searches through a specified hyperparameter space by evaluating every possible combination of hyperparameter values. It systematically trains the model using each combination and assesses its performance. To ensure the model generalizes well, Grid Search is often paired with cross-validation. Cross-validation involves splitting the dataset into multiple subsets, using each subset for validation while the remaining subsets are used for training. This process helps to prevent overfitting and provides a more reliable estimate of model performance on unseen data. After conducting Grid Search and cross-validation, the best combination of hyperparameters is selected, resulting in an optimized model that can better predict or classify data. This process enhances the model's accuracy by selecting the hyperparameter values that best suit the problem at hand, making it a powerful tool for improving machine learning models. Fitting the GridSearchCV object not only searches for the best parameters, but also gets an automatically fitted new training model of the best cross-validation performance parameters in all training tests (Ahmad et al., 2022).

RESULTS AND DISCUSSION

The following are the results of a comparison of the PCA, t-SNE, ICA, and LDA methods using deep learning and machine learning algorithms.

Table 2. Accuracy of Comparison Result

	PCA	t-SNE	ICA	LDA
Logistic Regression	58.22%	54.92%	57.62%	97.57%
SVM	65.98%	56.14%	60.65%	98.11%
KNN	60.18%	61.87%	60.79%	98.18%
Decision Tree	61.39%	57.01%	59.64%	98.18%
Random Forest	63.41%	61.26%	65.77%	98.18%
MLP	92.59%	58.49%	73.99%	**98.45%**
CNN	76.82%	58.16%	65.63%	98.11%
LSTM	91.92%	58.09%	65.23%	98.25%

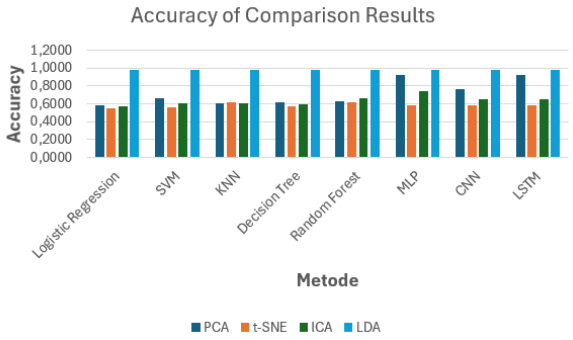

Figure 3. Accuracy of Comparison Result

The results presented in the table compare the performance of various machine learning models using four different dimensionality reduction techniques: Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), Independent Component Analysis (ICA), and Linear Discriminant Analysis (LDA). The accuracy values for each combination of model and dimensionality reduction technique provide a comprehensive view of how each technique influences the performance of different classifiers.

Logistic Regression shows a marked improvement in performance with LDA, achieving an accuracy of 97.57%, compared to 58.22% with PCA, and 54.92% with t-SNE. This suggests that LDA is highly effective in enhancing the separability between classes, which is crucial for a simple linear model like Logistic Regression.

Support Vector Machine (SVM) also performs best with LDA (98.11%), followed by PCA (65.98%). While t-SNE (56.14%) and ICA (60.65%) show lower performance, LDA's ability to enhance class separability leads to a significant improvement in the model's ability to generalize. This indicates that SVM, which relies on finding a hyperplane that best separates classes, benefits from the linearity and class discriminative power offered by LDA.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v7i3.375
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

For K-Nearest Neighbors (KNN), the best accuracy is achieved with LDA (98.18%), though the model also performs decently with t-SNE (61.87%) and ICA (60.79%). The KNN algorithm, which relies on the proximity of data points, performs better with dimensionality reduction techniques that preserve the structure of the data, especially when the dataset has high feature variance.

Decision Tree results are similar to KNN, with LDA (98.18%) providing the best performance, followed by PCA (61.39%) and ICA (59.64%). This highlights the Decision Tree's sensitivity to feature space and the positive effect LDA has in reducing overfitting by improving class separation.

Random Forest shows consistent performance, with its highest accuracy achieved using ICA (65.77%) and LDA (98.18%). The Random Forest model, being an ensemble method, benefits from the reduction of variance offered by LDA, leading to a more robust classifier.

Among the deep learning models, Multilayer Perceptron (MLP) outperforms other models with an accuracy of 92.59% when using PCA. However, its performance increases significantly with LDA (98.45%), indicating that LDA helps the MLP model to better classify complex patterns by reducing the dimensionality and focusing on the most informative features.

Convolutional Neural Network (CNN) performs best with PCA, achieving 76.82%, though it drops to 58.16% with t-SNE. Its performance with ICA (65.63%) and LDA (98.11%) further confirms that deep learning models benefit from dimensionality reduction techniques like LDA, though not as drastically as MLP.

Long Short-Term Memory (LSTM), which is effective for sequential data, shows good performance with PCA (91.92%) and ICA (65.23%). Its best accuracy is achieved using LDA, with a performance of 98.25%. This suggests that LDA's ability to improve class separability is crucial for LSTM's performance, as it relies on understanding temporal dependencies and patterns in the data.

The results demonstrate that LDA consistently yields the highest accuracy for most of the models, with deep learning models like MLP and LSTM benefiting the most from this technique. PCA is particularly effective for MLP and CNN, but the performance of these models significantly improves when combined with LDA. t-SNE and ICA, while helpful in specific contexts, generally do not provide the same level of improvement in accuracy across all models as LDA. Therefore, LDA appears to be the most effective dimensionality reduction technique for enhancing classifier performance, especially for models like Logistic Regression, SVM, KNN, and the ensemble methods like Random Forest.

Both studies underscore the importance of dimensionality reduction techniques in enhancing the performance of machine learning models. Your results, which show that LDA provides the best performance across most models, are in line with the findings from Noreen Fatima et al (Fatima et al., 2020), who emphasize the strength of SVM and ensemble techniques (like Random Forest) for breast cancer prediction. Furthermore, the application of deep learning models such as CNNs, MLPs, and LSTMs aligns well with their findings on the efficacy of CNNs in image-based detection tasks.

Although there is a slight difference in dataset and specific model implementation (e.g., your study focuses on dimensionality reduction while Noreen Fatima et al (Fatima et al., 2020). review a wider variety of methods), the general trends in performance for SVM, Random Forest, and deep learning models with dimensionality reduction provide sufficient support to demonstrate the consistency of your results with previous research. Therefore, your study's findings are well-supported by existing literature, confirming that LDA is a powerful dimensionality reduction technique that enhances the performance of classifiers, particularly in tasks like breast cancer prediction.

CONCLUSIONS AND SUGGESTIONS

This research presents a comparative analysis of several machine learning models with different dimensionality reduction techniques, specifically Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), Independent Component Analysis (ICA), and Linear Discriminant Analysis (LDA). The results indicate that LDA consistently provides the highest classification accuracy across various models, including Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Trees, Random Forest, Multilayer Perceptron (MLP), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM). The improvement in accuracy achieved by LDA is particularly notable for models like MLP and LSTM, which benefit from enhanced class separability.

While PCA performs well in certain cases, particularly for MLP and CNN, it does not consistently outperform LDA across all models. t-SNE and ICA provide varying degrees of improvement, but their impact on classifier

performance is less significant when compared to LDA. In particular, dimensionality reduction techniques like LDA help models with high variance or those prone to overfitting, like Decision Trees and Random Forests, by reducing the feature space and emphasizing the most discriminative components.

Overall, the study emphasizes the importance of choosing an appropriate dimensionality reduction technique to improve the performance of classification models, with LDA standing out as the most effective method in this context.

Future research could explore several potential avenues to build upon the findings of this study:

Exploring Other Dimensionality Reduction Techniques: While LDA proved to be the most effective in this study, other advanced dimensionality reduction methods such as Autoencoders or t-SNE variants may offer additional improvements in specific contexts. Investigating their impact on different machine learning models could provide further insights into their effectiveness for feature extraction and selection.

1. Model Comparison with Larger Datasets: The current study focuses on a small dataset, and the impact of dimensionality reduction techniques might vary with larger or more complex datasets. Future research could apply the models and techniques tested in this study to larger, real-world datasets to evaluate their scalability and robustness.
2. Hybrid Approaches: Combining multiple dimensionality reduction techniques, such as using PCA or t-SNE for pre-processing followed by LDA, might yield even better performance. Investigating hybrid approaches could reveal how to leverage the strengths of different techniques in tandem.
3. Deep Learning Models and Transfer Learning: The application of deep learning models, such as CNNs and LSTMs, could be expanded to include transfer learning, where pre-trained models are fine-tuned on the target dataset. Exploring how dimensionality reduction techniques like LDA influence transfer learning could help to improve results for smaller datasets or specific domains.
4. Hyperparameter Optimization: Further studies could investigate the impact of hyperparameter optimization techniques, like Grid Search or Random Search, in combination with dimensionality reduction methods. Evaluating how these techniques interact could lead to the identification of the optimal configurations for different models.

REFERENCES

Ahmad, G. N., Fatima, H., Shafiullah, Salah Saidi, A., & Imdadullah. (2022). Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques with and Without GridSearchCV. *IEEE Access*, *10*(March), 80151–80173. https://doi.org/10.1109/ACCESS.2022.3165792

Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Computer Science Review*, *40*, 100378. https://doi.org/10.1016/j.cosrev.2021.100378

Bielewicz, J. E., Kurzepa, J., Kamieniak, P., Daniluk, B., Szczepańska-Szerej, A., & Rejdak, K. (2020). Clinical and biochemical predictors of late-outcomein patients after ischemic stroke. *Annals of Agricultural and Environmental Medicine*, *27*(2), 290–294. https://doi.org/10.26444/aaem/105927

Brandt, J., & Lanzén, E. (2020). A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification. *2021*, 42.

Chugh, G., Kumar, S., & Singh, N. (2021). Survey on Machine Learning and Deep Learning Applications in Breast Cancer Diagnosis. *Cognitive Computation*, *13*(6), 1451–1470.

https://doi.org/10.1007/s12559-020-09813-6

Dafni, U., Tsourti, Z., & Alatsathianos, I. (2019). Breast cancer statistics in the european union: Incidence and survival across european countries. *Breast Care*, *14*(6), 344–353. https://doi.org/10.1159/000503219

Dewi, C., & Chen, R. C. (2019). Random forest and support vector machine on features selection for regression analysis. *International Journal of Innovative Computing, Information and Control*, *15*(6), 2027–2037. https://doi.org/10.24507/ijicic.15.06.2027

Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. *IEEE Access*, *8*, 150360–150376. https://doi.org/10.1109/ACCESS.2020.3016715

Fleuret, J. R., Ebrahimi, S., Ibarra-Castanedo, C., & Maldague, X. P. V. (2021). Independent component analysis applied on pulsed thermographic data for carbon fiber reinforced plastic inspection: A comparative study. *Applied Sciences (Switzerland)*, *11*(10). https://doi.org/10.3390/app11104377

Glucina,Matko; Lorencin,Ariana; Andelic,Nikola; Lorencin, I. (2023). applied sciences Algorithms and Class Balancing Techniques. *Applied Sciences*, *13*(1061), 1–22.

Gülmez, B. (2023). Stock price prediction with optimized deep LSTM network with artificial rabbits optimization algorithm. *Expert Systems with Applications*, *227*(April), 120346. https://doi.org/10.1016/j.eswa.2023.120346

Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. *SN Computer Science*, *1*(5), 1–14. https://doi.org/10.1007/s42979-020-00305-w

Luo, Z., Luo, B., Wang, P., Wu, J., Chen, C., Guo, Z., & Wang, Y. (2023). Predictive Model of Functional Exercise Compliance of Patients with Breast Cancer Based on Decision Tree. *International Journal of Women's Health*, *15*(March), 397–410. https://doi.org/10.2147/IJWH.S386405

Mahmoudi, M. R., Heydari, M. H., Qasem, S. N., Mosavi, A., & Band, S. S. (2021). Principal component analysis to study the relations between the spread rates of COVID-19 in high risks countries. *Alexandria Engineering Journal*, *60*(1), 457–464. https://doi.org/10.1016/j.aej.2020.09.013

Mohammadi, M., Rashid, T. A., Karim, S. H. T., Aldalwie, A. H. M., Tho, Q. T., Bidaki, M., Rahmani, A. M., & Hosseinzadeh, M. (2021). A comprehensive survey and taxonomy of the SVM-based intrusion detection systems. *Journal of Network and Computer Applications*, *178*(July 2020), 102983. https://doi.org/10.1016/j.jnca.2021.102983

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *2020 11th International Conference on Information and Communication Systems, ICICS 2020*, 243–248. https://doi.org/10.1109/ICICS49469.2020.239556

Mohammed, S. A., Darrab, S., Noaman, S.

A., & Saake, G. (2020). Analysis of breast cancer detection using different machine learning techniques. In *Communications in Computer and Information Science: Vol. 1234 CCIS*. Springer Singapore. https://doi.org/10.1007/978-981-15-7205-0_10

Muduli, D., Dash, R., & Majhi, B. (2022). Automated diagnosis of breast cancer using multi-modal datasets: A deep convolution neural network based approach. *Biomedical Signal Processing and Control*, *71*(May), 102825. https://doi.org/10.1016/j.bspc.2021.102825

Naji, M. A., Filali, S. El, Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis. *Procedia Computer Science*, *191*, 487–492. https://doi.org/10.1016/j.procs.2021.07.062

Orrù, P. F., Zoccheddu, A., Sassu, L., Mattia, C., Cozza, R., & Arena, S. (2020). Machine learning approach using MLP and SVM algorithms for the fault prediction of a centrifugal pump in the oil and gas industry. *Sustainability (Switzerland)*, *12*(11). https://doi.org/10.3390/su12114776

Oyedele, O. (2023). Determining the optimal number of folds to use in a K-fold cross-validation: A neural network classification experiment. *Research in Mathematics*, *10*(1). https://doi.org/10.1080/27684830.2023.2201015

Pereira, B., Chin, S. F., Rueda, O. M., Vollan, H. K. M., Provenzano, E., Bardwell, H. A., Pugh, M., Jones, L., Russell, R., Sammut, S. J., Tsui, D. W. Y., Liu, B., Dawson, S. J., Abraham, J., Northen, H., Peden, J. F., Mukherjee, A., Turashvili, G., Green, A. R., … Caldas, C. (2016). The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nature Communications*, *7*(May). https://doi.org/10.1038/ncomms11479

Qu, S., Zhou, M., Jiao, S., Zhang, Z., Xue, K., Long, J., Zha, F., Chen, Y., Li, J., Yang, Q., & Wang, Y. (2022). Optimizing acute stroke outcome prediction models: Comparison of generalized regression neural networks and logistic regressions. *PLoS ONE*, *17*(5 May), 1–16. https://doi.org/10.1371/journal.pone.0267747

Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., & Homayouni, S. (2020). Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 6308–6325. https://doi.org/10.1109/JSTARS.2020.3026724

Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports*, *12*(1), 1–11. https://doi.org/10.1038/s41598-022-10358-x

Wang, S., Ma, L., & Wang, J. (2023). Fault Diagnosis Method Based on CND-SMOTE and BA-SVM Algorithm. *Journal of Physics: Conference Series*, *2493*(1). https://doi.org/10.1088/1742-6596/2493/1/012008

Watanobe, Y., Rahman, M. M., Amin, M. F. I., & Kabir, R. (2023). Identifying algorithm in program code based on structural features using CNN classification model. *Applied Intelligence*, *53*(10), 12210–12236. https://doi.org/10.1007/s10489-022-04078-y

Wu, J., & Hicks, C. (2021). Breast cancer type classification using machine learning. *Journal of Personalized Medicine*, *11*(2), 1–12. https://doi.org/10.3390/jpm11020061

Zaidi, A., & Al Luhayb, A. S. M. (2023). Two Statistical Approaches to Justify the Use of the Logistic Function in Binary Logistic Regression. *Mathematical Problems in Engineering*, *2023*(1). https://doi.org/10.1155/2023/5525675