Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Integrating Hybrid Statistical and Unsupervised LSTM-Guided Feature Extraction for Breast Cancer Detection

De Rosal Ignatius Moses Setiadi [1,*], Arnold Adimabua Ojugo [2], Octara Pribadi [3], Etika Kartikadarma [1], Bimo Haryo Setyoko [4], Suyud Widiono [5], Robet [3], Tabitha Chukwudi Aghaunor [6], and Eferhire Valentine Ugbotu [7]

[1] Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia; e-mail : moses@dsn.dinus.ac.id; etika.kartikadarma@dsn.dinus.ac.id
[2] Department of Computer Science, Federal University of Petroleum Resources Effurun, Delta State 330102, Nigeria; e-mail : ojugo.arnold@fupre.edu.ng
[3] Department of Informatics Engineering, STMIK TIME, Medan 20212, Indonesia; e-mail : octarapribadi@stmik-time.ac.id; robet@stmik-time.ac.id
[4] Pusat Teknologi Informasi dan Pangkalan Data, Universitas Islam Negeri Salatiga, Salatiga 50716, Indonesia; e-mail : bimo_hs@uinsalatiga.ac.id
[5] Department of Computer Engineering, Faculty of Science and Technology, Universitas Teknologi Yogyakarta, Yogyakarta 55285, Indonesia; e-mail : suyud.w@uty.ac.id
[6] Department of Computer Science, School of Data Intelligence and Technology, Robert Morris University, Pittsburgh, PA 15108, United States; e-mail : cxast461@rmu.edu; tabitha.aghaunor@gmail.com
[7] Department of Data Science, College of Science and Engineering, University of Salford, Manchester M5 4WT, United Kingdom; e-mail : eferhire.ugbotu@gmail.com
***** Corresponding Author : De Rosal Ignatius Moses Setiadi

Abstract: Breast cancer is the most prevalent cancer among women worldwide, requiring early and accurate diagnosis to reduce mortality. This study proposes a hybrid classification pipeline that integrates Hybrid Statistical Feature Selection (HSFS) with unsupervised LSTM-guided feature extraction for breast cancer detection using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Initially, 20 features were selected using HSFS based on Mutual Information, Chi-square, and Pearson Correlation. To address class imbalance, the training set was balanced using the Synthetic Minority Oversampling Technique (SMOTE). Subsequently, an LSTM encoder extracted nonlinear latent features from the selected features. A fusion strategy was applied by concatenating the statistical and latent features, followed by re-selection of the top 30 features. The final classification was performed using a Support Vector Machine (SVM) with an RBF kernel and evaluated using 5-fold cross-validation and a held-out test set. The experimental results showed that the proposed method achieved an average training accuracy of 98.13%, an F1-score of 98.13%, and an AUC-ROC of 99.55%. On the held-out test set, the model reached an accuracy of 99.30%, a precision of 100%, and an F1-score of 99.05%, with an AUC-ROC of 0.9973. The proposed pipeline demonstrates improved generalization and interpretability compared with existing methods such as LightGBM-PSO, DHH-GRU, and ensemble deep networks. These results highlight the effectiveness of combining statistical selection and LSTM-based latent feature encoding in a balanced classification framework.

Keywords: Breast cancer detection; Ensemble feature selection; Feature fusion; Healthcare AI; Imbalance problem; Interpretable machine learning; Unsupervised LSTM.

Received: April, 4th 2025
Revised: May, 1st 2025
Accepted: May, 4th 2025
Published: May, 5th 2025

1. Introduction

Recent advances in artificial intelligence (AI) technology have brought about significant transformations in the medical world, especially in data-based and imaging-based disease diagnosis. AI enables faster and more accurate analysis processes and supports efficient clinical decision-making systems, including the detection of various types of cancer based on images and numerical data [1]–[4]. This trend strengthens AI's position as a potential solution to the increasing diagnostic workload and limited medical personnel, especially in radiology and oncology.

One of the main challenges in the field of oncology is breast cancer, which is the type of cancer with the highest incidence in women in the world. According to the World Health Organization, there are more than 2.3 million new cases and 670,000 deaths from breast cancer globally, making it the leading cause of cancer death in women in almost all countries[5]. Early and accurate diagnosis is a key strategy to increase the effectiveness of treatment and reduce mortality. In this context, various artificial intelligence approaches have been used to build decision support systems that can automatically classify cancer diagnosis results[6]–[11].

One of the benchmark datasets widely used in breast cancer classification studies is the Wisconsin Diagnostic Breast Cancer (WDBC). This dataset consists of 30 numerical features representing the statistical characteristics of cell nuclei from the results of fine needle aspiration (FNA) digitization. It has become the evaluation standard in many studies [12]–[14]. Zhu et al. [15] proposed an integration between SHAP-RF-RFE and LightGBM optimized using Particle Swarm Optimization (PSO). It achieved an accuracy of 99.0% and an AUC of 0.9870 on the testing data, demonstrating the superiority of the interpretability-based feature selection. Meanwhile, Natarajan et al. [16] applied the Dynamic Harris Hawks Optimization (DHH) approach combined with the gated recurrent unit (GRU) and performed feature engineering based on FFT and PCA. They reported an accuracy of 98.05% and an F1-score of 98.28% on the WDBC dataset. In another study, Sreehari and Dhinesh Babu [17] developed an Aggregated Coefficient Ranking-based Feature Selection (ACRFS) method combining Mutual Information (MI), Chi-square (χ^2), and Pearson Correlation Coefficient (PCC) for feature ranking. They tested it on various models, including a support vector machine (SVM) and a random forest (RF), and showed improved accuracy despite using a smaller feature subset. A study by Al Reshan et al. [18] proposed the integration of a Deep Neural Network (DNN) with a stacking ensemble approach (DNN-SEM) and used an extra tree classifier (ETC) for feature selection on four breast cancer datasets, including WDBC. This model achieved high accuracy, up to 99.62%, with a very competitive F1-score and MCC, demonstrating the superiority of deep learning-based ensemble models in breast cancer classification.

Although these methods show competitive results, each has its limitations. Deep learning models such as GRU and convolutional neural networks (CNN) can capture data complexity but require intensive training, large data volumes, and are prone to overfitting [19]. On the other hand, linear feature selection techniquessometimes can not maintain a rich representation of the original features, especially when dealing with nonlinear and highly correlated biological data. Even in the best approaches, such as Zhu et al. [15], the feature weighting process relies on supervised learning and the interpretability of ensemble models, but does not integrate features from non-linear representations obtained through sequential extraction, such as long short-term memory (LSTM).

In this situation, an approach that combines the strengths of conventional feature selection and nonlinear latent mapping is needed. Although originally developed for sequential data, LSTM has the capacity to represent dependencies between features in the latent space without the need for supervised training. Its ability as an unsupervised encoder makes it a potential candidate for more comprehensive feature extraction [20], [21], especially when combined with statistically based feature selection. In addition to modeling and feature selection approaches, another common challenge in medical datasets such as WDBC is the imbalanced class distribution, where the number of benign samples is often greater than that of malignant ones. This imbalance can cause the model to be biased toward the majority class [22]. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) approach is used, which synthetically adds new samples to the minority class based on interpolation from nearest neighbors [23], [24]. The use of SMOTE has been shown to be effective in improving the class distribution and increasing the sensitivity of the model in various cases such as studies [23], [25]–[27].

Based on these assumptions, an approach that integrates a combination of feature selection techniques and LSTM encoders is believed to be able to produce complementary feature representations, where the combination of feature selection techniques can filter the most statistically relevant features, and LSTM captures the non-linear latent structure of the data. This combined representation is then fused into one final feature vector, with the hypothesis that the combination of these two sources of information can improve the classification performance. SVM is used as a classifier because of its ability to handle high-dimensional data efficiently [28], as well as its stability, which has been proven in previous studies on the WDBC dataset. Based on this framework, this study contributes in several important aspects:

- Proposing an integration of feature selection techniques that we call HSFS and non-supervised LSTM encoders as a hybrid approach for feature selection and extraction;
- Developing a feature fusion mechanism to combine statistical and nonlinear information in one representation;
- Developing an efficient classification pipeline based on the SVM with comprehensive metric evaluation; and
- Comparing the performance of this approach to existing methods on the WDBC benchmark dataset.

This paper is organized into six main sections. Section 2 describes the preliminary chapters, covering the basic concepts and initial approaches, such as HSFS, unsupervised feature extraction with LSTM, feature fusion strategies, and related literature reviews. Section 3 explains the proposed method in detail, from preprocessing, feature selection, and data balancing with SMOTE to the classification process. Section 4 presents the experimental results, including the dataset description, model performance evaluation, and supporting visualizations. Section 5 compares the proposed method with another recent approach based on the same dataset, namely WDBC. Finally, Section 6 concludes the main findings and provides further development directions for further studies.

2. Preliminaries

2.1. Hybrid Statistical Feature Selection

Feature selection is an important process in machine learning that aims to select the most relevant and informative feature subset from the original data to improve model accuracy, reduce computational complexity, and prevent overfitting. Therefore, selecting the right features is crucial in building an efficient and accurate classification model[29], [30]. The ensemble feature selection approach is increasingly being used to improve the stability and effectiveness of the selection process. Ensemble feature selection combines the results of several selection methods to obtain a more representative and reliable feature subset than a single method. Combining feature selection methods has been used in several studies, such as [31]–[34].

This study adopts previous studies and proposes the Hybrid Statistical Feature Selection (HSFS) method. HSFS is a simple but effective customization strategy for feature selection by combining three independent statistical criteria: Mutual Information (MI), Chi-Square (χ^2), and Pearson Correlation Coefficient (PCC). These three methods were selected on the basis of their complementary evaluative characteristics: MI can capture non-linear relationships between features and labels, χ^2 is suitable for measuring the association of categorical features to classes, and PCC effectively identifies linear correlations. This combination aims to provide a more comprehensive evaluation of feature relevance while overcoming the limitations of each method when used alone.

Unlike some ensemble approaches that apply explicit weighting or parameter optimization, this method only uses the arithmetic mean of the ranking positions obtained from each metric. This approach is chosen to avoid one metric's dominance and maintain the neutrality of aggregation when the scales between metrics are not homogeneous. If direct score summation is used, the difference in numerical scales between metrics can cause significant bias in feature selection. In contrast, aggregation based on relative ranking is considered more stable, interpretable, and can be consistently reproduced, especially when applied to complex biomedical data with high correlations between features.

Mutual Information (MI) measures the level of dependence between a feature X and a label Y, including nonlinear relationships. The MI formula is defined in (1).

$$MI(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (1)$$

Chi-Square (χ^2) evaluates the strength of association between a feature and the target class based on the expected and observed frequencies, which is described in Equation (2).

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

On the other hand, PCC measures the strength and direction of the linear relationship between the features and target labels. The absolute value of PCC is used to avoid directional bias, which is shown in Equation (3).

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3}$$

After the scores of each metric are calculated, each feature f_i is assigned a ranking position $R_{MI}(f_i), R_{\chi^2}(f_i), R_{PCC}(f_i)$. The final ranking is calculated by the arithmetic mean of the three, as Equation (4).

$$R_{HSFS}(f_i) = \frac{R_{MI}(f_i) + R_{\chi^2}(f_i) + R_{PCC}(f_i)}{3} \tag{4}$$

The feature with the lowest average rank value is considered the most relevant and is selected for further classification. With this approach, HSFS can balance the sensitivity between metrics and provide more stable and reproducible feature selection results without the complexity of additional parameter tuning.

2.2. Unsupervised LSTM-Guided Feature Extraction

Long Short-Term Memory (LSTM) is a neural network architecture that is classically designed to handle sequential data and can retain long-term information through its internal memory mechanism [35], [36]. Although commonly used in domains such as natural language processing and time signals, the internal structure of LSTM allows its application to model dependencies between features in tabular data, including biomedical data, without labels or supervised training [37].

This study uses LSTM as an unsupervised encoder to extract denser dimensional latent feature representations. The goal is to map high-dimensional inputs into a more structured and informative nonlinear latent space to complement the statistically selected features. Unlike conventional deep learning training, this extraction process is carried out only through a forward pass, without an explicit training stage, without a loss function, and without a backpropagation process.

Technically, the input feature selection results using HSFS are in the form of a matrix $X \in \mathbb{R}^{n \times d}$, transformed into a three-dimensional form $X' \in \mathbb{R}^{n \times 1 \times d}$ to be compatible with the LSTM architecture. The encoding process can be formulated in Equation (5) and (6).

$$h_t = LSTM(X'_t), \tag{5}$$

$$z = ReLU(W.h_t + b) \tag{6}$$

Where h_t is the hidden output of LSTM, W is the weight matrix of the dense layer, and is the bias vector. The activation result z is the feature vector of the dense layer activation result as the final representation.

The advantage of this approach lies in its ability to capture nonlinear latent structures that may not be reached by conventional statistical selection methods such as MI, χ^2, or PCC. In complex breast cancer data, features may interact in nontrivial patterns not sufficiently explained by linear correlation or categorical association alone.

2.3. Feature Fusion Strategy

The fusion of various feature sources into a single unified representation is an approach widely used in developing modern deep learning models, both for tabular, image, and multimodal data [38]–[40]. This strategy combines the representational power of various features, such as statistical, spatial, and temporal features. Thus, the classification model can obtain richer and complementary information [41], [42]. In the context of biomedical data, feature fusion is also a promising approach to combine domain knowledge (through explicit features) with latent representations (through feature embedding).).

This study uses a strategy to combine the results ofthe statisticall feature selection based on HSFS and the results of nonlinear feature extraction from the LSTM encoder. The goal is to form a combined feature vector that reflects the statistical relevance to the label and captures the latent dependencies between features not covered by conventional statistical methods.

Technically, the HSFS selected feature is denoted as $X_{HSFS} \in \mathbb{R}^{n \times d_1}$, and the latent representation from the LSTM encoder as $Z_{LSTM} \in \mathbb{R}^{n \times d_2}$. These two matrices are combined horizontally in concatenation to produce a combined feature vector (see Equation (7)).

$$X_{fusion} = [X_{HSFS} \| Z_{LSTM}] \in \mathbb{R}^{n \times (d_1 + d_2)} \tag{7}$$

This fusion is performed before the classification stage, assuming that this combined representation has a betterbetter discriminative capacity. To maintain classification efficiency and prevent redundancy, the HSFS-based feature selection process is reapplied to X_{fusion}, so that only the most informative combined features are used in the final classification stage. This fusion strategy not only strengthens the data representation but also maintains the interpretability and efficiency of the model and supports the integration of statistical and representation learning approaches in a single unified classification pipeline.

2.4. Related works

Several approaches for automatic breast cancer detection using the Wisconsin dataset have been developed. Oyelade et al. [12] introduced the ST-ONCODIAG system, which combines an ontology-based semantic reasoning model with a Select-and-Test approach. This system demonstrated its effectiveness in extracting and representing medical knowledge through domain-based logical rule inference. On the other hand, Alshayeji et al. [13] evaluated the performance of artificial neural networks (ANNs) without feature selection. They found that a simple ANN model was effective for breast cancer classification, emphasizing that complex structures are not always necessary for optimal performance. In addition, Raza et al. [14] proposed an explainable artificial intelligence (XAI) approach based on ensemble learning. They explored the combination of bagging, boosting, and stacking techniques with classification algorithms such as RF, XGBoost, and LightGBM.

Al Reshan et al.[18] developed a Deep Neural Network-based Stacking Ensemble Model (DNN-SEM) that combines deep learning techniques with an ensemble approach for breast cancer classification. This model combines two level-1 models (DBN and ANN) with four level-0 models (XGBoost, Logistic Regression, RF, and SVM) and applies feature selection using an Extra Tree Classifier (ETC). Although their study tested four different datasets, one of which was WDBC. The results showed that the deep learning-based ensemble approach can significantly improve the classification accuracy and offer better interpretability than conventional methods.

In addition, Sreehari and Babu [17] proposed the Aggregated Coefficient Ranking-based Feature Selection (ACRFS) method, which is a feature selection strategy based on the aggregate ranking of three problem solvers: Mutual Information (MI), Pearson Correlation Coefficient (PCC), and Chi-square. This study used the WDBC dataset along with three other datasets. ACRFS aims to identify important features more stably and efficiently by combining the advantages of the three selection techniques. Experimental results show that this approach significantly improves the accuracy and efficiency of ML models such as RF, KNN, and SGD. Specifically, the SGD model experienced an accuracy improvement of up to 19%, with substantial improvements in the F1-score and MCC after applying ACRFS, while the Decision Tree method showed a performance degradation.

Zhu et al. [15] proposed the SHAP-RF-RFE algorithm, which integrates Shapley values (SHAP) in a RF-based Recursive Feature Elimination (RFE) framework. The best classification model used was LightGBM, which was combined with Particle Swarm Optimization (PSO) for hyperparameter optimization. As a result, the accuracy reached 99.0%, with a precision of 100%, a recall of 97.4%, an F1-score of 98.68%, and an AUC of 0.9870. In addition, 26 selected features were determined as the most relevant features, with radius_worst, area_worst, and perimeter_worst being the most impactful features.

In addition, Alhassan et al. [16] proposed a GRU-based deep learning approach that is intelligently optimized using the Dynamic Harris Hawks Optimization (DHHO) metaheuristic algorithm. Instead of relying on classical feature selection, they directly tune the GRU

parameters through an intelligent search based on exploration and exploitation. The results show that the dynamically optimized GRU achieves higher classification accuracy than the baseline GRU and other ML methods.

However, most previous approaches still have limitations in integrating statistical-based feature selection and nonlinear feature extraction from deep learning models, which can complement each other to improve generalization. In addition, only a few studies have combined aggregation-based feature selection methods with latent feature extraction using unsupervised deep learning. They are then combined in the final feature selection stage before classification. Therefore, this study proposes a hybrid approach based on HSFS-LSTM fusion that more efficiently represents important features while maintaining a high discriminative ability for breast cancer classification.

3. Proposed Method

The proposed method is a hybrid approach that combines statistical feature selection (HSFS) with nonlinear feature extraction using unsupervised LSTM, followed by feature fusion and classification using SVM. The process starts with data preprocessing using Min-Max normalization, then HSFS is applied, which calculates MI, χ^2, and PCC scores, then sorts the features based on the average of the three ranks.

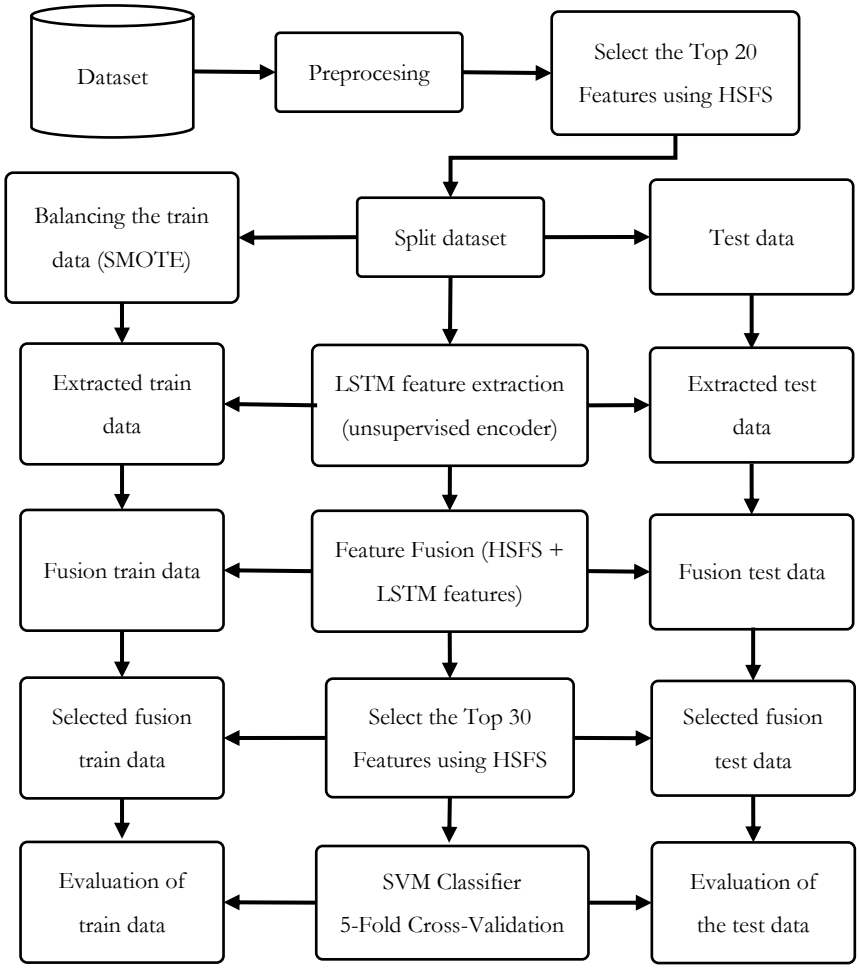

Figure 1. Proposed method illustration flow.

Next, the top 20 features are extracted, and the data are trained using SMOTE to address class imbalance. An LSTM encoder is used for the unsupervised extraction of latent features from the numerical data, and the features are fused with the original features to form a combined representation. Then, the feature selection process is repeated on the fused data using the HSFS method to select the top 30 features. Finally, the SVM classification model was

trained and evaluated using 5-fold stratified cross-validation and testing on the testing data to measure accuracy, precision, recall, F1, specificity, confusion matrix, Kappa, and MCC.

Table 1. Details of the proposed step-by-step.

Step	Description	Configuration/Function
Preprocessing	The dataset was loaded and normalized using MinMaxScaler. The diagnosis label is encoded in numerical format.	MinMaxScaler(), LabelEncoder()
Feature Selection (HSFS)	Feature selection is based on the average ranking of the HSFS.	mutual_info_classif, chi2, pearsonr, select the top 20 features
Splitting and balancing	Data are split into training and testing sets with a 75:25 ratio. SMOTE is applied to balance the class distribution.	train_test_split(test_size=0.25, stratify=y), SMOTE(random_state=42)
LSTM feature extraction	An unsupervised LSTM encoder extracts latent features from a numerical input with a 1×20 dimension.	LSTM(units=32), Dense(units=16, activation='relu')
Feature Fusion	The selected features from HSFS and latent features from LSTM are concatenated into a unified feature vector.	np.concatenate(axis=1)
Feature Re-selection	The selected features from the HSFS and latent features from the LSTM are concatenated into a unified feature vector.	HSFS to get top 30 features
SVM Classification	An SVM model was trained using 5-fold Stratified Cross-Validation and evaluated on the test set.	SVC(kernel='rbf', probability=True, random_state=42), StratifiedKFold(n_splits=5)
Evaluation	Model performance was evaluated using accuracy, precision, recall, F1-score, Kappa, MCC, AUC-ROC, confusion matrix, and specificity.	accuracy_score, precision_score, recall_score, f1_score, roc_auc_score, cohen_kappa_score, matthews_corrcoef, ConfusionMatrixDisplay

4. Results and Discussion

4.1. Dataset Description

This study used the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, accessed at https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic. The dataset contains digitized results of microscopic images from fine-needle aspiration (FNA) of breast cancer, with a total of 569 samples and 32 columns, consisting of 1 ID column, 1 diagnosis column (target: M = Malignant, B = Benign), and 30 numerical feature columns.

4.1.1 Feature Types and Descriptions

The features used in this dataset can be categorized into three numerical groups based on the calculation phase:

- Mean features (_mean) – Measures the average value of morphology-related properties such as radius_mean, texture_mean, perimeter_mean, and area_mean.
- Standard deviation features (_se) – Measures each feature's error or standard deviation across measurements.
- Worst value features (_worst) – Maximum value recorded during a measurement for each feature.

4.1.2 Diagnosis

The diagnosis label has been numerically encoded, with 1 for Malignant and 0 for Benign. The proportion of malignant cases was 37.3%, and the proportion of benign cases was

62.7%. Out of the 569 samples in the WDBC dataset, this corresponds to 212 malignant and 357 benign cases. This class distribution indicates a moderate imbalance, potentially affecting model training and performance if not adequately addressed.

4.1.3 Feature Correlation and Heatmap Visualization

Correlation analysis is important in understanding the relationship between dataset features and breast cancer diagnosis. In this study, the Pearson correlation coefficient was used, which results in values ranging from -1 to 1, with the interpretation as follows: values approaching 1 indicate a strong positive correlation, values near -1 indicate a strong negative correlation, and values around 0 indicate no linear correlation. Figure 2 shows the heatmap of the correlations among all features in the WDBC dataset. Red indicates a strong positive correlation between features, while blue indicates a weak or negative correlation.

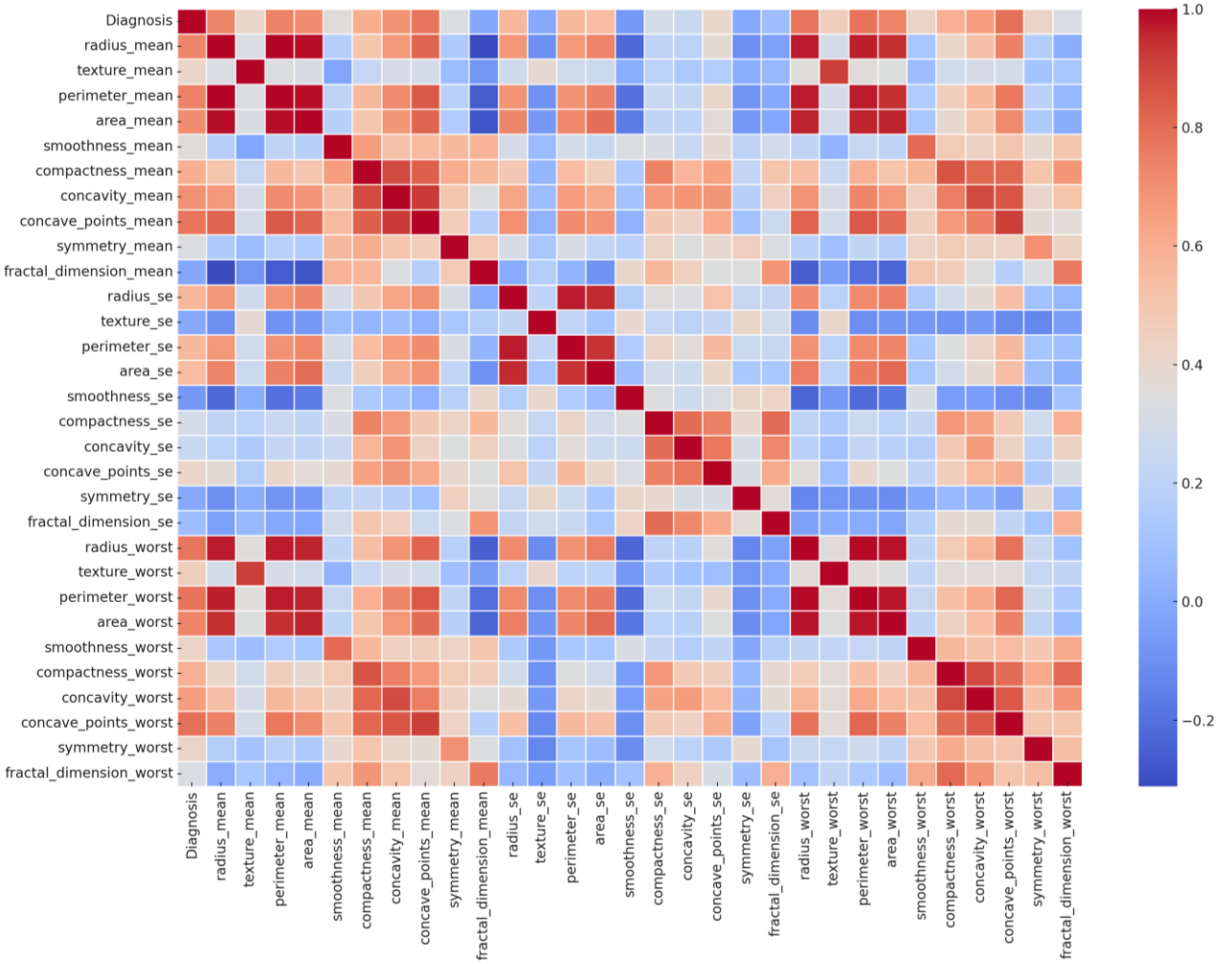

Figure 2. Correlation Heatmap of the WDBC Features

Some key findings from the heatmap visualization: First, the diagnosis variable highly correlates with features like radius_mean, perimeter_mean, area_mean, concavity_mean, concave_points_mean, and compactness_mean. This indicates that higher values in these features tend to correlate with the malignant class. Second, features from the "_worst" group, such as radius_worst, perimeter_worst, and area_worst, also showed a strong correlation, reflecting the structural consistency among the morphology-related features.

Meanwhile, the "_se" group (standard error), such as texture_se, smoothness_se, and symmetry_se, tended to show lower correlation with diagnosis and among themselves, indicating high variability. These findings support using an HSFS feature selection method since highly correlated features may carry redundant information. Therefore, identifying features that are strongly correlated with the label but not with each other is crucial to maintaining model efficiency.

4.2. Preprocessing and HSFS Feature Selection

The dataset obtained from the UCI Wisconsin Breast Cancer Diagnostic Repository underwent several preprocessing steps. First, the categorical target variable diagnosis was encoded into binary labels (0 for benign, 1 for malignant) using the LabelEncoder. All numeric features were normalized into the range [0, 1] using MinMaxScaler to ensure uniform scaling across dimensions and to optimize model convergence. Next, HSFS was performed using three statistical criteria. Each feature was ranked individually based on each criterion, and an aggregate rank was calculated by averaging the ranks across all three metrics. The top 20 features with the highest aggregate scores were selected for further analysis and are presented in Table 2.

Table 2. Top 20 initial HSFS features.

Feature	MI	Chi2	PCC	Aggregate
symmetry_worst	0.094862	5.560093	0.416294	19.333333
texture_mean	0.094736	6.394071	0.415185	19
smoothness_worst	0.103943	5.675733	0.421465	18.333333
concave_points_se	0.126605	5.781996	0.408042	18
texture_worst	0.116552	8.741628	0.456903	16.333333
perimeter_se	0.272321	16.044344	0.556141	13.666667
radius_se	0.248513	17.324128	0.567134	13.333333
compactness_mean	0.211762	20.353176	0.596534	12.666667
area_se	0.341687	19.676975	0.548236	12.666667
compactness_worst	0.225653	20.992541	0.590998	12.333333
concavity_worst	0.314751	31.563031	0.65961	9.333333
radius_mean	0.366222	24.897293	0.730029	8.333333
area_mean	0.360925	29.328594	0.708984	8.333333
perimeter_mean	0.404168	26.528902	0.742636	6.666667
concavity_mean	0.37473	46.186395	0.69636	6.333333
radius_worst	0.453728	34.124937	0.776454	4.333333
area_worst	0.465511	35.043882	0.733825	4
concave_points_mean	0.439231	52.405743	0.776614	2.666667
perimeter_worst	0.47833	34.438091	0.782914	2.666667
concave_points_worst	0.436537	46.341648	0.793566	2.666667

4.3. Splitting and Balancing

After feature selection of the top 20 features using the HSFS method, the data were split into training and testing sets with a 75:25 ratio, using the stratify parameter to maintain a balanced class proportion between the training and testing data. However, the training data still showed a class imbalance, with 267 benign samples (0) and only 159 malignant samples (1). To address this issue, the SMOTE method was applied to generate synthetic data for the minority class through interpolation between the nearest neighbors. After oversampling with SMOTE, the number of samples in both classes became balanced, each with 267 samples. This process helps improve the classification model performance by avoiding bias toward the majority class. Figure 3 illustrates the class distribution before and after SMOTE balancing.

4.4 LSTM Feature Extraction, Feature Fusion, and Re-selection

After the initial feature selection stage, the process continues with feature extraction using LSTM in an unsupervised encoder architecture. Although the dataset is tabular, LSTM is designed to capture temporal patterns from the data. The features extracted by LSTM are then fused with the 20 selected features using HSFS. This process results in a new high-dimensional representation that combines the descriptive power of the classical features and the latent representation from the LSTM. Subsequently, a re-selection of features is performed on the fused feature set using the HSFS method again. From this second selection stage, the top 30 features with the highest aggregate scores were selected and used as the main

input for classification. Table 3 presents the list of selected top features based on aggregate scores before and after the fusion process.

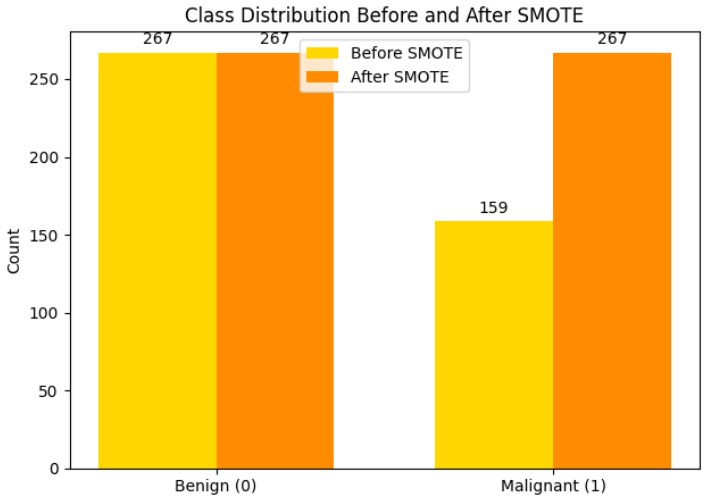

Figure 3. Dataset distribution before and after SMOTE.

Table 3. Top 30 features after the fusion process.

Feature	MI	Chi2	PCC	Aggregate
perimeter_se	0.294321	14.038611	0.542817	27
perimeter_se_lstm	0.294695	14.038611	0.542817	26.666667
radius_se_lstm	0.260151	15.475315	0.557076	26.333333
radius_se	0.261399	15.475315	0.557076	26
compactness_worst_lstm	0.222018	17.359451	0.564583	25.666667
compactness_worst	0.22258	17.359451	0.564583	25.333333
area_se_lstm	0.374756	16.57068	0.534224	25
area_se	0.375932	16.57068	0.534224	24.666667
compactness_mean_lstm	0.245307	18.713144	0.590653	23.666667
compactness_mean	0.246626	18.713144	0.590653	23.333333
concavity_worst_lstm	0.339845	26.073655	0.645003	19
concavity_worst	0.340001	26.073655	0.645003	18.666667
area_mean_lstm	0.404231	27.500479	0.688486	15
radius_mean	0.415331	24.041247	0.724212	15
radius_mean_lstm	0.415662	24.041247	0.724212	14.666667
area_mean	0.404307	27.500479	0.688486	14.666667
concavity_mean_lstm	0.391498	38.455948	0.666591	13.666667
concavity_mean	0.392801	38.455948	0.666591	13.333333
perimeter_mean_lstm	0.422419	25.366642	0.73446	13
perimeter_mean	0.424167	25.366642	0.73446	12.666667
area_worst	0.499705	32.402183	0.713863	9.666667
area_worst_lstm	0.500547	32.402183	0.713863	9.333333
radius_worst_lstm	0.497292	32.647572	0.769044	6.333333
radius_worst	0.499031	32.647572	0.769044	6
concave_points_mean	0.466715	46.603761	0.755062	5.5
concave_points_mean_lstm	0.466715	46.603761	0.755062	5.5
perimeter_worst	0.50817	32.618952	0.772561	5
concave_points_worst	0.450813	41.388151	0.789057	5
perimeter_worst_lstm	0.50867	32.618952	0.772561	4.666667
concave_points_worst_lstm	0.451875	41.388151	0.789057	4.666667

It should be noted that there are changes in the MI, χ^2, PCC, and aggregate scores for several features that reappear in the re-selection process. For example, the perimeter_se feature has an aggregate score of 19.33 in the initial selection but increases to 27.00 in the re-selection stage (see Table 3). This change is due to the differences in the evaluation context. The initial selection was performed only on the classical features, while the re-selection was performed on a combination of the classical and LSTM features. Because the MI, χ^2, and PCC metrics are relative to the feature set used, the order and weight of each feature may change in the context of feature fusion.

4.5. SVM Classification and Evaluation

After the feature extraction and re-selection process on the fused classical and LSTM features, the classification model was developed using SVM with an RBF kernel. To comprehensively evaluate model performance and prevent overfitting, a 5-fold cross-validation was conducted on the training data obtained from the SMOTE-balanced data, and a final evaluation was performed on a held-out test set that had never been seen during training.

The SVM performance results during the 5-fold cross-validation on the training data are presented in Table 4. The average accuracy reached 98.13%, with a precision of 98.53%, recall of 97.77%, F1-score of 98.13%, and AUC-ROC of 99.55%. The high Kappa and MCC scores indicate that the model accurately and consistently handles the minority class after the balancing process. The specificity reached 98.50%, indicating that the model effectively recognizes benign samples.

Table 4. Evaluation results of the proposed method.

Metrics	5-Fold Cross-Validation (Training)					Held-Out Test Set
	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	
Accuracy	0.9906	0.9813	0.9720	0.9626	1.0000	0.9930
Precision	1.0000	0.9636	1.0000	0.9630	1.0000	1.0000
Recall	0.9811	1.0000	0.9444	0.9630	1.0000	0.9811
Specificity	1.0000	0.9630	1.0000	0.9623	1.0000	1.0000
F1 Score	0.9905	0.9815	0.9714	0.9630	1.0000	0.9905
Kappa Score	0.9813	0.9626	0.9439	0.9252	1.0000	0.9850
MCC	0.9815	0.9633	0.9454	0.9252	1.0000	0.9851
AUC-ROC	0.9993	0.9996	0.9832	0.9951	1.0000	0.9973

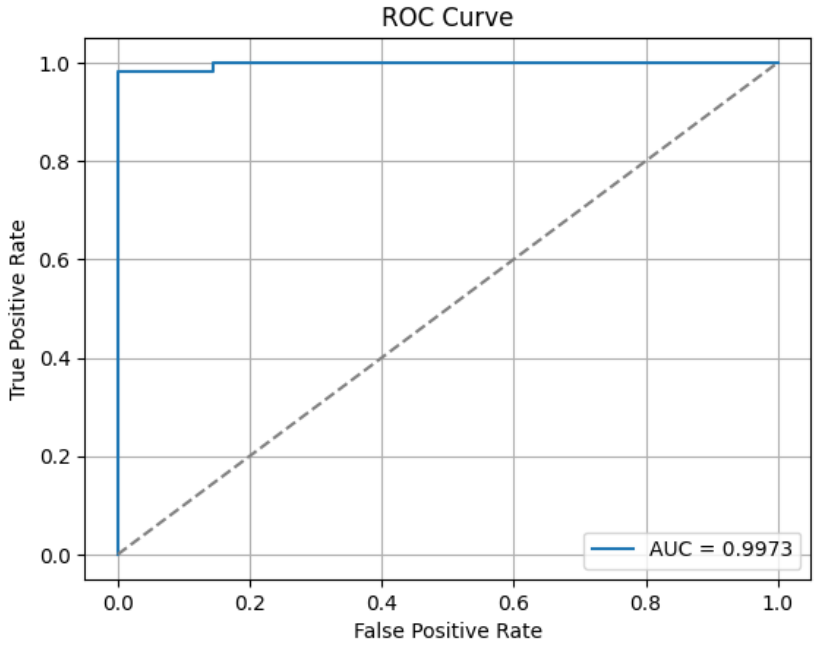

Figure 4. ROC curve results of the held-out test set.

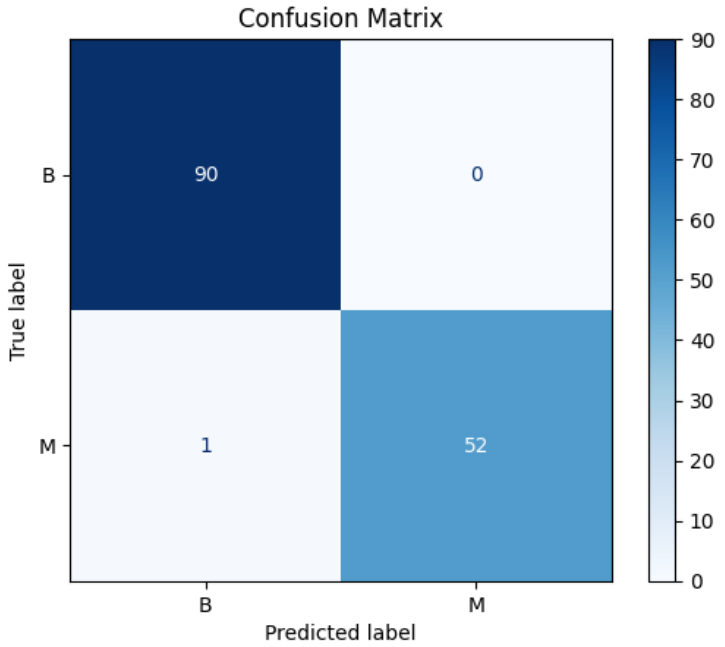

Figure 5. Confusion matrix of the held-out test set.

The held-out test set comprising 25% of the original dataset was evaluated to measure the model's generalization ability on unseen data. The test results showed an accuracy of 99.30%, precision of 100%, recall of 98.11%, and F1-score of 99.05%. The AUC value reached 0.9973, indicating an almost perfect separation between the benign and malignant cases. Specificity reached 100%, meaning no benign sample was misclassified. The confusion matrix in Figure 4 and the ROC curve in Figure 5 further support these findings.

The ROC curve in Figure 4 shows an AUC-ROC of 0.9973, which is close to perfect. AUC is an important metric because it reflects the model's ability to distinguish between positive and negative classes at various decision thresholds. The closer the value is to 1.0, the better the model's discriminatory performance. In addition, the confusion matrix in Figure 5 confirms that there was only one misclassification out of 143 samples, and there were no false positives. Specificity reached 100%, meaning no noncancer cases were misclassified as cancer, which is important to avoid overdiagnosis. Overall, combining LSTM-based feature extraction, fusion with classical features, and HSFS-based feature selection significantly enhanced the breast cancer classification performance on both the training data and the held-out test set.

5. Comparison

To evaluate the effectiveness of the proposed model, we conducted a comparative analysis against four recent studies that also used the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The comparison is focused on the held-out test set performance, as it provides a more realistic indication of the model's generalization ability. Table 5 summarizes the comparison of the key metrics across the models.

Based on Table 5, the proposed HSFS-LSTM-SVM approach outperforms the other models across most evaluation metrics on the held-out test set. Notably, it achieved the highest accuracy (99.30%), precision (100%), and specificity (100%), indicating superior generalization and low false positive rates—critical in medical diagnosis. Compared to LightGBM-PSO [15], which also attained high performance, our model improved the AUC-ROC by over 1%, demonstrating stronger class separation. Similarly, DHH-GRU [16] achieved slightly lower accuracy (98.05%) and F1-score (98.28%) than our method. Although SEM-DBN and SEM-ANN [18] leverage deep learning ensembles, their slightly lower metrics and lack of specificity reporting make them less robust. Importantly, our model maintains high performance without requiring complex stacking architectures or supervised deep training. The integration of SMOTE also contributes significantly by addressing class imbalance, which is

often neglected in prior works. Our method offers a practical interpretable and effective solution for breast cancer classification.

Table 5. Evaluation results of the proposed method.

Metrics	LightGBM-PSO [15]	DHH-GRU [16]	ACRFS-RF [17]	SEM-DBN [18]	SEM-ANN [18]	HSFS-LSTM-SVM (Ours)
Accuracy	0.9900	0.9805	0.9790	0.9872	0.9852	0.9930
Precision	1.0000	0.9809	0.9763	0.9880	0.9855	1.0000
Recall	0.9740	0.9815	0.9795	0.9802	0.9797	0.9811
Specificity	-	-	-	0.9990	0.9890	1.0000
F1 Score	0.9868	0.9828	0.9779	0.9900	0.9880	0.9905
Kappa Score	-	-	0.9558	-	-	0.9850
MCC	-	-	-	0.9880	0.9860	0.9851
AUC-ROC	0.9870	-	0.9559	-	-	0.9973

6. Conclusions

This study presented a hybrid approach combining HSFS and unsupervised LSTM-guided feature extraction to improve breast cancer classification using the WDBC dataset. The model achieved a high discriminative capability by fusing statistically selected features and latent representations and re-applying feature selection on the fused space. Using SMOTE during training successfully mitigated the class imbalance, enhancing the recall and F1-score. The final SVM classification with the RBF kernel achieved 99.30% accuracy and perfect precision (100%) on a held-out test set. Comparative analysis with state-of-the-art models demonstrates our method's superior generalization and balance. The study contributes a lightweight yet effective framework that avoids complex training while offering interpretability and high performance. Future work may involve extending this hybrid strategy to multiclass or multimodal medical datasets and testing alternative fusion or dimensionality reduction strategies to further improve efficiency and robustness.

Author Contributions: Conceptualization: DRIMS and AAO.; Methodology: DRIMS; Software: OP and EK; Validation: AAO, OP, EK, BHS, SW, R, TCA, and EVU.; Formal analysis: EK.; Investigation: AAO, BHS, SW, R, TCA, and EVU; Resources: DRIMS; Data curation: OP, BHS, SW, and R; Writing—original draft preparation: DRIMS; Writing—review and editing: AAO, OP, EK, BHS, SW, R, TCA, and EVU.; Visualization: X.X.; Supervision: X.X.; Project administration: EK; funding acquisition: All. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset used in this study can be accessed publicly at the link https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] D. S. Stamoulis and C. Papachristopoulou, "Artificial Intelligence in Radiology, Emergency, and Remote Healthcare: A Snapshot of Present and Future Applications," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 3, pp. 228–234, Oct. 2024, doi: 10.62411/faith.3048-3719-38.

[2] O. Okolo, B. Y. Baha, and M. D. Philemon, "Using Causal Graph Model variable selection for BERT models Prediction of Patient Survival in a Clinical Text Discharge Dataset," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 4, pp. 455–473, Mar. 2025, doi: 10.62411/faith.3048-3719-61.

[3] M. B. Teferi and L. A. Akinyemi, "Deep Learning-Based Cross-Cancer Morphological Analysis: Identifying Histopathological Patterns in Breast and Lung Cancer," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 3, pp. 235–248, Oct. 2024, doi: 10.62411/faith.3048-3719-36.

[4] O. Jaiyeoba, O. Jaiyeoba, E. Ogbuju, and F. Oladipo, "AI-Based Detection Techniques for Skin Diseases: A Review of Recent Methods, Datasets, Metrics, and Challenges," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 3, pp. 318–336, Dec. 2024, doi: 10.62411/faith.3048-3719-46.

[5] World Health Organization (WHO), "Breast cancer," *who.int*, 2024. https://www.who.int/news-room/fact-sheets/detail/breast-cancer (accessed Apr. 30, 2025).
[6] S. Fanijo, "AI4CRC: A Deep Learning Approach Towards Preventing Colorectal Cancer," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 2, pp. 143–159, Sep. 2024, doi: 10.62411/faith.2024-28.
[7] K. Swanson, E. Wu, A. Zhang, A. A. Alizadeh, and J. Zou, "From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment," *Cell*, vol. 186, no. 8, pp. 1772–1791, Apr. 2023, doi: 10.1016/j.cell.2023.01.035.
[8] A. Yaqoob, R. Musheer Aziz, and N. K. Verma, "Applications and Techniques of Machine Learning in Cancer Classification: A Systematic Review," *Human-Centric Intell. Syst.*, vol. 3, no. 4, pp. 588–615, Sep. 2023, doi: 10.1007/s44230-023-00041-3.
[9] G. Sruthi, C. L. Ram, M. K. Sai, B. P. Singh, N. Majhotra, and N. Sharma, "Cancer Prediction using Machine Learning," in *2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM)*, Feb. 2022, pp. 217–221. doi: 10.1109/ICIPTM54933.2022.9754059.
[10] Y. Kumar *et al.*, "Automating cancer diagnosis using advanced deep learning techniques for multi-cancer image classification," *Sci. Rep.*, vol. 14, no. 1, p. 25006, Oct. 2024, doi: 10.1038/s41598-024-75876-2.
[11] A. Khalid *et al.*, "Breast Cancer Detection and Prevention Using Machine Learning," *Diagnostics*, vol. 13, no. 19, p. 3113, Oct. 2023, doi: 10.3390/diagnostics13193113.
[12] O. N. Oyelade, A. A. Obiniyi, S. B. Junaidu, and S. A. Adewuyi, "ST-ONCODIAG: A semantic rule-base approach to diagnosing breast cancer base on Wisconsin datasets," *Informatics Med. Unlocked*, vol. 10, no. December 2017, pp. 117–125, 2018, doi: 10.1016/j.imu.2017.12.008.
[13] M. H. Alshayeji, H. Ellethy, S. Abed, and R. Gupta, "Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach," *Biomed. Signal Process. Control*, vol. 71, no. PA, p. 103141, Jan. 2022, doi: 10.1016/j.bspc.2021.103141.
[14] A. D. Raha *et al.*, "Modeling and Predictive Analytics of Breast Cancer Using Ensemble Learning Techniques: An Explainable Artificial Intelligence Approach," *Comput. Mater. Contin.*, vol. 81, no. 3, pp. 4033–4048, 2024, doi: 10.32604/cmc.2024.057415.
[15] J. Zhu *et al.*, "An integrated approach of feature selection and machine learning for early detection of breast cancer," *Sci. Rep.*, vol. 15, no. 1, p. 13015, Apr. 2025, doi: 10.1038/s41598-025-97685-x.
[16] R. Natarajan, S. Krishna, H. L. Gururaj, F. Flammini, B. S. Alfurhood, and C. M. N. Kumar, "A Novel Hybrid Dynamic Harris Hawks Optimized Gated Recurrent Unit Approach for Breast Cancer Prediction," *Int. J. Comput. Intell. Syst.*, vol. 18, no. 1, p. 7, Jan. 2025, doi: 10.1007/s44196-024-00712-4.
[17] E. Sreehari and L. D. Dhinesh Babu, "A novel aggregated coefficient ranking based feature selection strategy for enhancing the diagnosis of breast cancer classification using machine learning," *Sci. Rep.*, vol. 15, no. 1, p. 4171, Feb. 2025, doi: 10.1038/s41598-025-87826-7.
[18] M. S. Al Reshan *et al.*, "Advanced breast cancer prediction using Deep Neural Networks integrated with ensemble models," *Chemom. Intell. Lab. Syst.*, vol. 262, no. January, p. 105399, Jul. 2025, doi: 10.1016/j.chemolab.2025.105399.
[19] P. S. R. C. Murty *et al.*, "Integrative hybrid deep learning for enhanced breast cancer diagnosis: leveraging the Wisconsin Breast Cancer Database and the CBIS-DDSM dataset," *Sci. Rep.*, vol. 14, no. 1, p. 26287, Nov. 2024, doi: 10.1038/s41598-024-74305-8.
[20] A. Sagheer and M. Kotb, "Unsupervised Pre-training of a Deep LSTM-based Stacked Autoencoder for Multivariate Time Series Forecasting Problems," *Sci. Rep.*, vol. 9, no. 1, p. 19038, Dec. 2019, doi: 10.1038/s41598-019-55320-6.
[21] L. Annamalai, V. Ramanathan, and C. S. Thakur, "Event-LSTM: An Unsupervised and Asynchronous Learning-Based Representation for Event-Based Data," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 4678–4685, Apr. 2022, doi: 10.1109/LRA.2022.3151426.
[22] S. Aymaz, "Unlocking the power of optimized data balancing ratios: a new frontier in tackling imbalanced datasets," *J. Supercomput.*, vol. 81, no. 2, p. 443, Jan. 2025, doi: 10.1007/s11227-025-06919-2.
[23] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on expansion and classification of imbalanced data based on SMOTE algorithm," *Sci. Reports 2021 111*, vol. 11, no. 1, pp. 1–11, 2021, doi: 10.1038/s41598-021-03430-5.
[24] S. M. Imran and A. Geetha, "Evaluating the Effectiveness of Smote for Imbalanced Data Expansion and Its Impact on Classification Accuracy," in *2024 First International Conference for Women in Computing (InCoWoCo)*, Nov. 2024, pp. 1–7. doi: 10.1109/InCoWoCo64194.2024.10863344.
[25] F. O. Aghware *et al.*, "Enhancing the Random Forest Model via Synthetic Minority Oversampling Technique for Credit-Card Fraud Detection," *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 407–420, Mar. 2024, doi: 10.62411/jcta.10323.
[26] D. R. I. M. Setiadi, K. Nugroho, A. R. Muslikh, S. W. Irianandа, and A. A. Ojugo, "Integrating SMOTE-Tomek and Fusion Learning with XGBoost Meta-Learner for Robust Diabetes Recognition," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 23–38, May 2024, doi: 10.62411/faith.2024-11.
[27] C. C. Odiakaose *et al.*, "Hypertension Detection via Tree-Based Stack Ensemble with SMOTE-Tomek Data Balance and XGBoost Meta-Learner," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 3, pp. 269–283, Dec. 2024, doi: 10.62411/faith.3048-3719-43.
[28] F. S. Gomiasti, W. Warto, E. Kartikadarma, J. Gondohanindijo, and D. R. I. M. Setiadi, "Enhancing Lung Cancer Classification Effectiveness Through Hyperparameter-Tuned Support Vector Machine," *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 396–406, Mar. 2024, doi: 10.62411/jcta.10106.
[29] J. A. Ingio, A. S. Nsang, and A. Iorliam, "Optimizing Rice Production Forecasting Through Integrating Multiple Linear Regression with Recursive Feature Elimination," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 2, pp. 96–108, Aug. 2024, doi: 10.62411/faith.2024-17.
[30] D. R. I. M. Setiadi, S. Widiono, A. N. Safriandono, and S. Budi, "Phishing Website Detection Using Bidirectional Gated Recurrent Unit Model and Feature Selection," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 2, pp. 75–83, Jul. 2024, doi: 10.62411/faith.2024-15.
[31] M. I. Akazue, I. A. Debekeme, A. E. Edje, C. Asuai, and U. J. Osame, "UNMASKING FRAUDSTERS: Ensemble Features Selection to Enhance Random Forest Fraud Detection," *J. Comput. Theor. Appl.*, vol. 1, no. 2, pp. 201–211, Dec. 2023, doi: 10.33633/jcta.v1i2.9462.

[32] K. Natarajan, D. Baskaran, and S. Kamalanathan, "An adaptive ensemble feature selection technique for model-agnostic diabetes prediction," *Sci. Rep.*, vol. 15, no. 1, p. 6907, Feb. 2025, doi: 10.1038/s41598-025-91282-8.
[33] A. Hashemi, M. B. Dowlatshahi, and H. Nezamabadi-pour, "Ensemble of feature selection algorithms: a multi-criteria decision-making approach," *Int. J. Mach. Learn. Cybern.*, vol. 13, no. 1, pp. 49–69, Jan. 2022, doi: 10.1007/s13042-021-01347-z.
[34] D. P. M. Abellana, R. R. Roxas, D. M. Lao, P. E. Mayol, and S. Lee, "Ensemble Feature Selection in Binary Machine Learning Classification: A Novel Application of the Evaluation Based on Distance from Average Solution (EDAS) Method," *Math. Probl. Eng.*, vol. 2022, pp. 1–13, Sep. 2022, doi: 10.1155/2022/4126536.
[35] W. Xia, W. Zhu, B. Liao, M. Chen, L. Cai, and L. Huang, "Novel architecture for long short-term memory used in question classification," *Neurocomputing*, vol. 299, pp. 20–31, Jul. 2018, doi: 10.1016/j.neucom.2018.03.020.
[36] S. Adamu, A. Iorliam, and Ö. Asilkan, "Exploring Explainability in Multi-Category Electronic Markets: A Comparison of Machine Learning and Deep Learning Approaches," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 4, pp. 440–454, Mar. 2025, doi: 10.62411/faith.3048-3719-58.
[37] S. M. Al-Selwi *et al.*, "RNN-LSTM: From applications to modeling techniques and beyond—Systematic review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 5, p. 102068, Jun. 2024, doi: 10.1016/j.jksuci.2024.102068.
[38] L. R. Zuama, D. R. I. M. Setiadi, A. Susanto, S. Santosa, H.-S. Gan, and A. A. Ojugo, "High-Performance Face Spoofing Detection using Feature Fusion of FaceNet and Tuned DenseNet201," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 4, pp. 385–400, Feb. 2025, doi: 10.62411/faith.3048-3719-62.
[39] A. Pathirana *et al.*, "A Reinforcement Learning-Based Approach for Promoting Mental Health Using Multimodal Emotion Recognition," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 2, pp. 124–142, Sep. 2024, doi: 10.62411/faith.2024-22.
[40] N. R. Pratama, D. R. I. M. Setiadi, I. Harkespan, and A. A. Ojugo, "Feature Fusion with Albumentation for Enhancing Monkeypox Detection Using Deep Learning Models," *J. Comput. Theor. Appl.*, vol. 2, no. 3, pp. 427–440, Feb. 2025, doi: 10.62411/jcta.12255.
[41] Z. Golrizkhatami and A. Acan, "ECG classification using three-level fusion of different feature descriptors," *Expert Syst. Appl.*, vol. 114, pp. 54–64, Dec. 2018, doi: 10.1016/j.eswa.2018.07.030.
[42] Y. Qiao *et al.*, "A multi-modal fusion model with enhanced feature representation for chronic kidney disease progression prediction," *Brief. Bioinform.*, vol. 26, no. 1, Nov. 2024, doi: 10.1093/bib/bbaf003.