AUIQ Technical Engineering Science

Manuscript 1041

Classify Cancers Types Using miRNAs Profiles and Machine Learning

Wafaa Khazaal Shams

AUIQ TECHNICAL ENGINEERING SCIENCE 2025;2:167–175

ORIGINAL STUDY

Classify Cancers Types Using miRNAs Profiles and Machine Learning

Wafaa Khazaal Shams

Center of Information Technology, University of Mustansiriyah, Baghdad, 14022, Iraq

ABSTRACT

Cancer classification is important for early diagnosis and effective treatment. This study proposes a hybrid framework that combines Fisher Discriminant analysis (FD) with Linear Support Vector Machine (LSVM) and Long Short-Term Memory (LSTM) networks. MicroRNAs (miRNAs) expression profiles for three types of cancer are used to evaluate the model. These are: lung-adenocarcinoma, ovary and pancreas cancers. The classification process is done using 5-fold cross valuation. Results show significant performance for LSTM compared to LSVM with and without features reduction. However, using FD analysis results in high accuracy reached 0.97and 0.96 for LSTM and LSVM respectively, using 50 miRNAs. The accuracy of all models improved with increased feature number sets. Another test has been done to find the relation between these types of cancers based on the classification process. Most pancreas cancer samples are classified as lung-adenocarcinoma cancer compared to ovary cancer in an unseen test. Overall, the results demonstrate the effects of LSTM-FD as a dependable tool for cancer diagnosis. Further, our model detects original cancer like the ovary and pancreas accurately.

Keywords: Cancer classification, Fisher discriminate, miRNA profile, Machine learning

1. Introduction

Detecting cancer at its earliest stages is crucial for enabling effective treatment and improving survival rates [1, 2]. Despite the availability of various technologies and methods for identifying early symptoms and tissue changes, many cancer cases are still diagnosed at advanced stages [2, 3]. MicroRNAs (miRNAs) are small, non-coding RNA molecules that play a significant role in regulating gene expression [4]. Studies have shown that miRNA expression is often dysregulated in cancer cells. This abnormal expression can result from several factors, including gene amplification, abnormal transcriptional control, epigenetic alterations, and defects in miRNA biogenesis machinery [4–6]. Such dysregulation has been observed in various types of cancer, including breast, lung, colorectal, and prostate cancers [4, 6]. MicroRNAs expression profiling, particularly through liquid biopsy, offers a noninvasive and tissue-specific method that holds promise for cancer screening and diagnosis. Additionally, miRNA expression patterns can provide insights into prognosis and treatment response [4, 5]. Machine learning (ML) techniques have been increasingly applied to analyze miRNA expression profiles for cancer classification and prognosis prediction. These computational approaches enable fast and accurate identification of miRNA-related diseases [8, 9]. However, several challenges remain, particularly the imbalance between the small number of available samples and the high dimensionality of miRNA data. Machine learning (ML) methods have been used for miRNA expression profiles in cancer classification and prognosis prediction. Using computational methods help for fast and effective predication of the miRNA related disease [8, 9]. However, several challenges arise when applying machine learning (ML) techniques to miRNA data. A major issue is the imbalance between the small number of available samples and the high

Received 1 June 2025; revised 4 July 2025; accepted 14 July 2025.
Available online 13 August 2025

E-mail address: dr_wafaa@uomustansiriyah.edu.iq (W. K. Shams).

https://doi.org/10.70645/3078-3437.1041

dimensionality of the data, which often leads to overfitting. Additionally, false negatives may occur if low-abundance miRNAs are not accurately detected [10, 11]. Therefore, careful preparation and preprocessing of miRNA profiles are crucial to ensure data quality and reliability. Feature selection tools have shown substantial improvements in ML performance, particularly when applied to datasets from the Gene Expression Omnibus (GEO). These tools help reduce the dimensionality of gene expression data, optimize feature sets, and rank miRNAs based on their ability to distinguish between cancerous and non-cancerous tissues, as well as among different cancer types [12–14]. Support Vector Machines (SVMs) are widely used and robust classifiers in the medical field. They are particularly effective for small sample sizes due to their strong generalization capabilities and stability [8, 12, 13, 15, 16]. More recently, deep learning models have shown promising performance in cancer detection tasks. Among them, Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network (RNN) that have demonstrated an ability to learn long-range dependencies in sequential data[17]. Several studies have utilized LSTM networks for various genomic applications. Tasdelen and Sen [18] applied LSTM to classify pre-miRNA vs. pseudo-miRNA, achieving an accuracy above 0.94, although their focus was not on disease detection. Aburass et al. [19] employed a hybrid CNN-LSTM model to classify DNA mutation sequences across five cancer types, achieving a recall of approximately 80%. Ali et al. [20] applied LSTM to miRNA expression data and reported an accuracy of up to 0.94 in detecting kidney cancer types. These findings suggest that LSTM networks can effectively model complex patterns in miRNA expression, even though miRNA data are not inherently sequential. In this study, we propose a hybrid framework combining Fisher Discriminant Analysis (FDA) [21] with LSTM networks to detect lung adenocarcinoma, ovarian, and pancreatic cancers. We compare the performance of our model with Linear SVM (LSVM) [22]. Our investigation highlights the ability of LSTM networks to capture ordered, interdependent relationships—characteristics often present in genomic expression data. A notable limitation of this study is the small sample size (100 samples), which may affect generalizability and model robustness.

2. Material and methods

2.1. Data description and Fisher discriminate score

The microarray dataset used in this study originates from the work of Søkilde et al. [23]. It is publicly available through the National Library of Medicine's Gene Expression Omnibus (GEO) database [24]. The dataset includes miRNA expression profiles for 15 distinct cancer types. Specifically, 208 formalin-fixed, paraffin-embedded (FFPE) tissue samples from human subjects were profiled using a locked nucleic acid (LNA)-enhanced microarray platform [23]. For the purpose of this study, we focused on three cancer types: lung adenocarcinoma (25 samples), ovarian cancer (33 samples), and pancreatic cancer (42 samples). Each sample includes expression data for 1887 unique miRNA identifiers (IDs). The first step in our classification pipeline is the application of a feature selection method to optimize the miRNA dataset. This process reduces the dimensionality of the dataset and removes irrelevant or redundant features, thereby improving classification performance. We employed Fisher Discriminant Analysis (FDA), which is based on the statistical properties of features to identify those that best differentiate between classes. The Fisher score for the multi-class setting (three cancer types) is defined as shown in Eq. (1) [14]:

$$FD(j) = \frac{\sum_{i=1}^{C} N_i \left(\bar{x}_{i,j} - \mu_j\right)}{\sum_{i=1}^{C} N_i \sigma^2(x_{i,j})} \tag{1}$$

Where C is a number of classes, N_i is a number of samples in each class, μ_i is a mean of feature j, $\bar{x}_{i,j}$ is mean of feature j in class i, $x_{i,j}$ is the sample vector of class i and $\sigma^2(x_{i,j})$ variance of feature j in class i.

2.2. Classification methods

For the classification process, we utilized two classifiers: Linear Support Vector Machine (LSVM) and Long Short-Term Memory (LSTM) networks. LSVM is a binary classifier that separates data into two classes by finding a hyperplane that maximizes the margin between them. For multi-class classification with three classes, LSVM employs multiple binary models, each distinguishing between a pair of classes. These are then combined using strategies such as one-vs-one or one-vs-all [25]. In this study, the one-vs-one approach was utilized. Classifier parameters were tuned using 10-fold cross-validation. For LSVM, the regularization parameter CCC was set to 0.1, standardization was enabled, the dual formulation was used, and the loss function was set to squared hinge.

On the other hand, Long Short-Term Memory (LSTM) networks are a specialized type of Recurrent Neural Network (RNN) designed to capture complex dependencies across inputs through their gated architecture. Traditional RNNs face limitations such as vanishing and exploding gradients, which hinder their ability to learn long-range interactions.

Table 1. Networks architecture.

Layer No.	LSTM Architecture
1	Input Layer
2	LSTM Layer (sequence output)
3	Dropout
4	LSTM Layer (last output)
5	Dropout
6	Fully Connected
7	Softmax
8	Classification Layer

Table 2. Networks hyperparameters.

Hyperparameter	LSTM Selection
Optimizer	'adam'
MaxEpochs	100
MiniBatchSize	20
InitialLearnRate	0.01
GradientThreshold	0.1
L2Regularization	0.01

LSTMs address these issues using memory cells and three gating mechanisms—input, forget, and output gates—that regulate the flow of information and allow the model to preserve relevant dependencies [17, 26]. In this study, miRNAs are represented as a set of input features rather than sequences. However, these features may exhibit inherent dependencies due to biological co-regulation and shared functional pathways. The LSTM network is employed to capture these underlying relationships among miRNAs, even in the absence of explicit sequential ordering. This approach aims to leverage the LSTM's strength in modeling contextual dependencies to improve prediction performance. The architectural details of LSTM models are presented in Table 1. To prevent overfitting, L2 regularization ($\lambda = 0.01$) and dropout (rate = 0.2) were applied after each hidden layer. Each hidden layer consists of 20 nodes for LSTM model. Hyperparameters were optimized using a 10-fold cross-validation combined with manual grid search, as summarized in Table 2.

A 5-fold cross-validation procedure was applied to evaluate the classification model using different numbers of selected features, starting from 5, 10, 20, 50, 100, 150, 200, 300, 500, 1000, and 1887 features. For each fold, fold-wise Z-score normalization was performed on the features by computing the mean and standard deviation from the training data only. These statistics were then applied to normalize both the training and test sets to avoid data leakage. The normalized training data was used to train the model using the train Network function. The same process was repeated using the full feature set without any feature selection to evaluate the impact of feature selection on model performance. Additionally, to investigate the relationships and separability between classes, a further test was conducted where the model was trained on data from two classes and tested on the third class. To assess the classification performance comprehensively, four metrics were calculated: accuracy, precision, recall, and F1-score, defined by the following equations [27]:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{2}$$

$$Percision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1\text{-}score = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{5}$$

Though *TP* is a number of true positive, *TN* is a number of *true* negative, *FP* is a number of false positive, and *FN* is a number of false negative. Precision indicates the false alarms when it has low value means cancer type is classified wrongly when it is not belonged to that type of cancer. However high value means cancer types are correctly classified. Recall concerns on detection the actual types therefore low recall means missing cancer cases that need to treatment. High recall means truly indicates of cancer cases. F1-score indicates how well classifier model performs for each specific cancer. Accuracy assesses the performance of model to identify cancers classes correctly [27]. All experiments and analyses were conducted using MATLAB R2021b.

3. Result

3.1. Generalization and overfitting analysis

The classification results are presented in this section. To evaluate the performance and generalization ability of the classifiers, we first assessed overfitting and generalization behavior [28]. The dataset was divided into an 80% training set and a 20% external test set. A 5-fold cross-validation was then performed on the training portion to tune and evaluate the model performance. After training on the full training set, the final models were validated on the unseen external test set to confirm their generalization ability. Table 3 shows the cross-validation results and external test results. Clearly, the model achieved high and balanced performance across all three classes. During cross-validation, the model showed strong and balanced results across the three classes, with

Table 3. Cross validation and external test result using 100 features set with LSTM.

	Cross-Validation Results			External Test Results		
classes	Precision	Recall	F1-score	Precision	Recall	F1-score
1	0.88	0.79	0.83	1.00	1.00	1.00
2	0.88	0.85	0.87	0.75	1.00	0.86
3	0.89	0.97	0.93	1.00	0.75	0.86
average	0.89	0.87	0.88	0.92	0.92	0.90

Table 4. Cross validation and external test result using 100 features set with LSVM.

	Cross-Validation Results			External Test Results		
classes	Precision	Recall	F1-score	Precision	Recall	F1-score
1	0.93	0.90	0.91	0.83	1.00	0.91
2	0.94	0.88	0.89	1.00	0.86	0.92
3	0.92	0.94	0.92	1.00	1.00	1.00
average	0.93	0.91	0.91	0.94	0.95	0.94

F1-scores ranging from 0.83 to 0.93, and an average accuracy of approximately 88–90%. These results indicated effective learning without signs of overfitting during training. The final evaluation of the external test set—data completely unseen during training and tuning—confirmed this generalization. The model achieved 90% accuracy, with perfect precision and recall for one class and strong F1 scores (0.857) for the remaining two. The close alignment between cross-validation and test results demonstrates that the model generalized well to new data and did not overfit the training set.

Table 4 shows the cross-validation and external test results for the LSVM model. The cross-validation demonstrates consistently high precision, recall, and F1-scores across all three classes (approximately 0.9 or above). This indicates that the model performs well and stably on the training data, with balanced sensitivity (recall) and precision—correctly identifying true positives while minimizing false positives. On the independent external test set (20% hold-out), the model achieves an overall accuracy of 95%, confirming strong generalization to unseen data. Although precision and recall vary slightly between classes, the F1 scores remain very high (close to or equal to 1 for classes 2 and 3), demonstrating that the model maintains robust predictive performance beyond the training environment. Results show no overfitting.

3.2. Classification performance based on number of feathers set

Fig. 1 illustrates the average classification accuracy for detecting lung adenocarcinoma, ovarian, and pancreatic cancers using different classifiers. The classification was performed using 5-fold cross-validation. The gray and red lines in the figure represent the accuracy results of LSVM and LSTM, respectively, using the full miRNA expression feature set without any feature selection. In contrast, the yellow and black lines correspond to the accuracy of LSVM and LSTM, respectively, when applying Fisher Discriminant (FD) feature selection. Notably, classification accuracy improved with FD. The highest accuracy was achieved by LSTM, reaching 0.97 and 0.96 with 50 and 100 selected features, respectively. LSVM achieved 0.96 accuracy with 50 and 300 selected features. Detailed classification results are summarized in Table 3. Both classifiers showed improved performance with FD-selected features, although differences between them were relatively small. Overall, classifier performance increased sharply with the first 50 selected features and then stabilized in the range of 100 to 500 features.

To determine whether there is a significant difference in accuracy among the four classifiers across various feature sizes, we applied the Friedman test

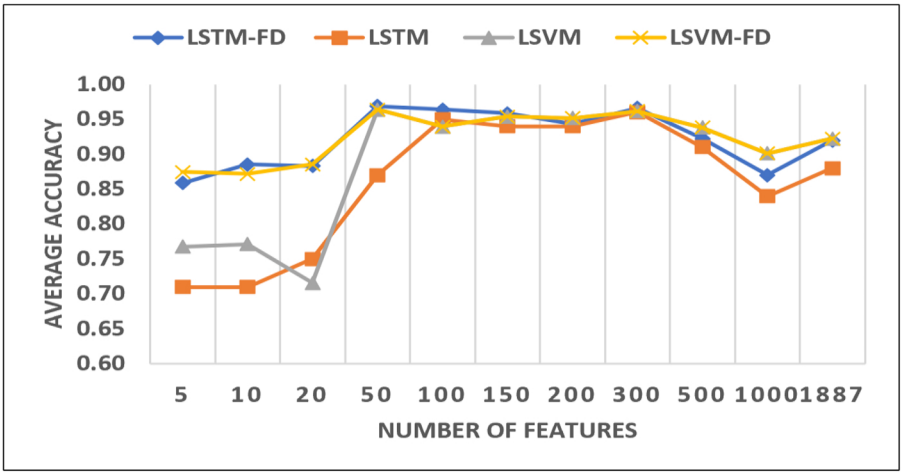

Fig. 1. Average accuracy in respect to number of features.

Table 5. Average accuracy and standard deviation for classification methods.

Features	LSTM-FD	LSTM	LSVM-FD	LSVM
5	0.86 ± 2.5	0.71 ± 3.5	0.87 ± 4.2	0.77 ± 3.6
10	0.89 ± 1.3	0.71 ± 4.8	0.87 ± 3.3	0.77 ± 4.5
20	0.88 ± 2.8	0.75 ± 4.5	0.89 ± 4.3	0.72 ± 2.6
50	0.97 ± 3.5	0.87 ± 5	0.96 ± 3.5	0.88 ± 5
100	0.96 ± 4.3	0.95 ± 1.5	0.94 ± 2.9	0.94 ± 2.8
150	0.96 ± 1.8	0.94 ± 3.5	0.95 ± 2.5	0.96 ± 3.6
200	0.94 ± 4.6	0.94 ± 3.6	0.95 ± 3.3	0.94 ± 4.2
300	0.97 ± 2.7	0.96 ± 4.5	0.96 ± 2.7	0.94 ± 3.5
500	0.92 ± 4.2	0.91 ± 2.7	0.94 ± 3.0	0.90 ± 3.3
1000	0.87 ± 3.8	0.84 ± 3.9	0.90 ± 2.5	0.93 ± 2.9
1887	0.92 ± 3.6	0.88 ± 4.5	0.92 ± 3.4	0.92 ± 4.2

[29]. This non-parametric test is appropriate for comparing more than two related groups. In this case, the classifiers use repeated measurements from the same datasets (i.e., feature sets across different sizes). The Friedman test works by ranking the classifiers for each run and evaluating whether the differences in average ranks are statistically significant. If the Friedman test indicates a significant difference, we proceed with post hoc pairwise comparisons to identify which specific classifiers differ. Table 5 presents the results of these tests. The results show that both the LSTM-FD and LSVM-FD classifiers significantly outperform the baseline LSTM classifier. However, there is no significant difference between LSTM-FD and LSVM-FD, nor between LSVM and either LSTM or LSVM-FD. When we narrow our analysis to the subset of feature sizes ranging from 20 to 300 features, we find no statistically significant difference among the classifiers ($p \geq 0.05$). This suggests that within this optimal feature size range, the classifiers perform comparably.

Fig. 2 shows the classification metric for detecting lung-adenocarcinoma cancer using 100 miRNA features for all methods. The classification results highlight the strong potential of miRNA sequences in identifying cancer types. For lung adenocarcinoma,

Table 6. Friedman test results.

Pair	p-value	Interpretation
LSTM-FD vs LSTM	0.0044	**Significant difference** ($p < 0.05$) — LSTM-FD outperforms LSTM
LSTM-FD vs LSVM-FD	1.0000	No significant difference
LSTM-FD vs LSVM	1.0000	No significant difference
LSVM-FD vs LSTM	0.0112	**Significant difference** — LSVM-FD outperforms LSTM
LSVM vs LSTM	0.2460	No significant difference
LSVM-FD vs LSVM	1.0000	No significant difference

the LSTM classifier achieved perfect recall (1.0), while LSVM reached 0.96. This suggests that miRNA expression carries significant discriminatory information for detecting lung adenocarcinoma compared to other cancer types. Fig. 3 presents the classification metrics for detecting ovarian cancer. The LSTM with feature decomposition (FD) demonstrated excellent performance, with a precision of 0.97 and a recall of 1.0. The lowest performance was LSVM without FD, which yielded a precision of 0.80 and a recall of 0.85. For pancreatic cancer, similarly high precision was observed. LSTM with FD achieved a precision of 1.0, while LSVM with FD achieved 0.97. The recall scores were 0.95 for LSTM with FD and 0.91 for LSVM without FD, as shown in Fig. 4. Overall, feature decomposition (FD) had a positive impact on both types of classifiers. Although the performance difference with and without FD was relatively small for LSTM, the effect was more pronounced for LSVM, particularly when using a smaller number of features. It is important to note that the reported results were based on using 100 features, and the benefit of FD becomes more evident when the number of features is lower. Notably, LSVM outperformed other models in detecting lung adenocarcinoma, suggesting its suitability for this specific cancer type. Across all classifiers and cancer types, the F1 scores remained consistently close to 1.0, indicating a high degree of

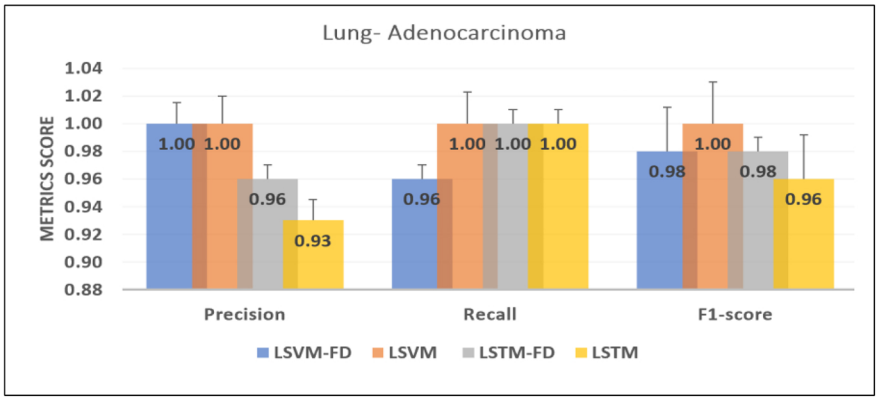

Fig. 2. Classification metric for detection lung-adenocarcinoma.

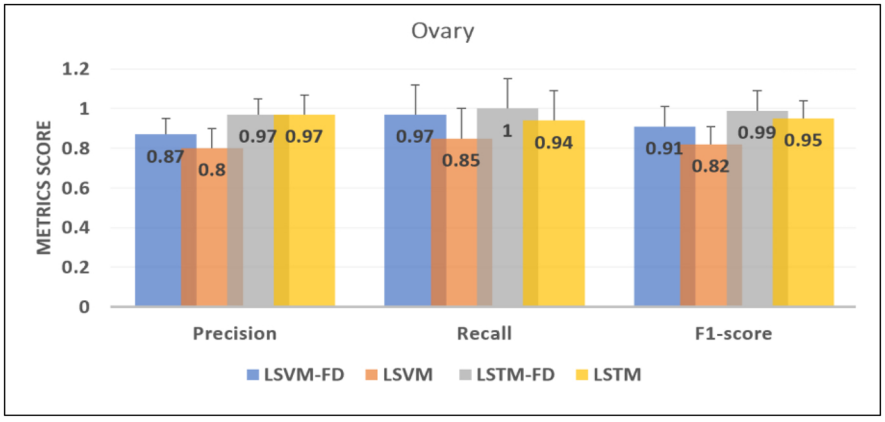

Fig. 3. Classification metric for ovary cancer.

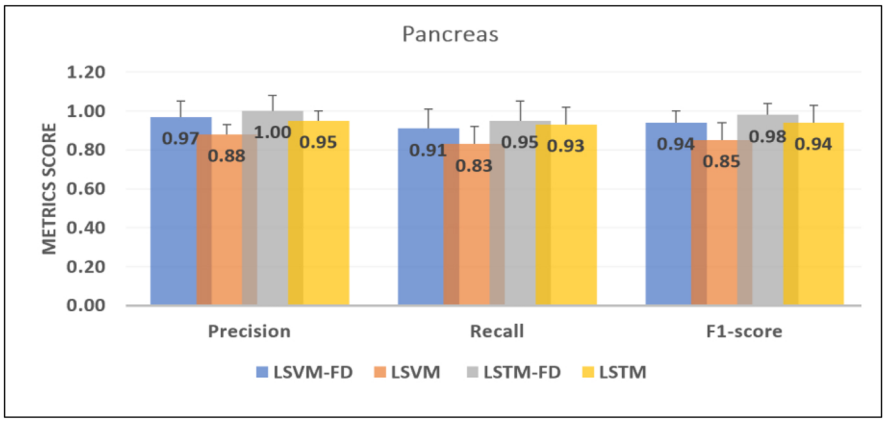

Fig. 4. Classification metric of pancreas cancer.

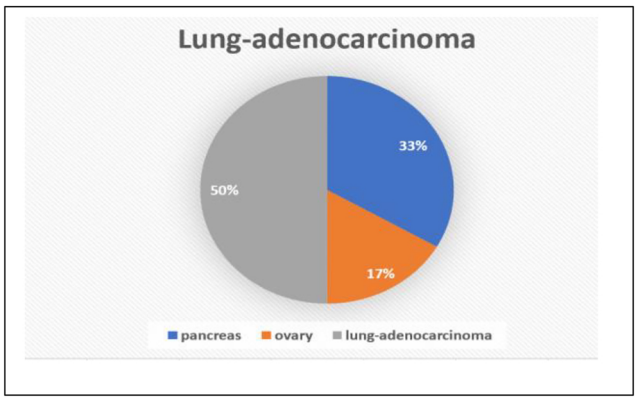

Fig. 5. Percentage miscalssifed lung cancer.

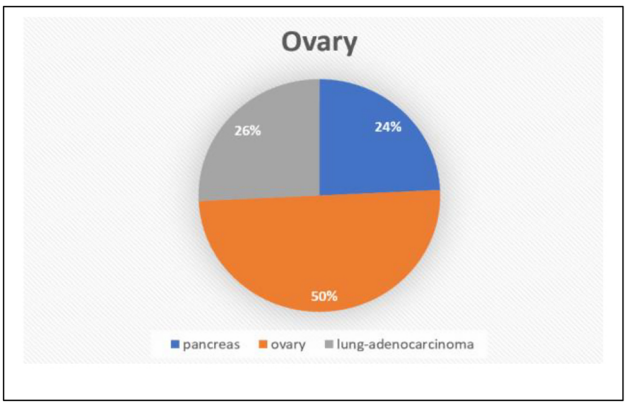

Fig. 6. Percentage miscalssifed ovary cancer.

reliability and a strong balance between precision and recall.

3.3. Unseen test for each class

Figs. 5 to 7 present the results of testing each cancer type separately using LSVM. Fig. 5 shows the percentage of misclassified samples among the 25 lung adenocarcinoma cases. Notably, 33% of the samples were misclassified as pancreatic cancer, while 17% were classified as ovarian cancer. For ovarian cancer (Fig. 6), the misclassifications were more evenly distributed: 26% of the samples were misclassified as lung adenocarcinoma and 24% as pancreatic

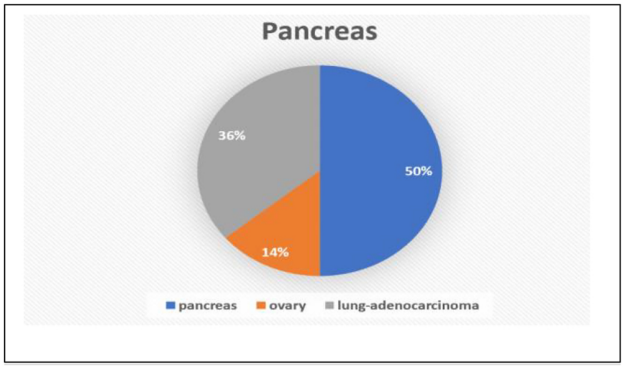

Fig. 7. Percentage miscalssifed pancreas cancer.

cancer. Fig. 7 illustrates the misclassification rates for pancreatic cancer. In this case, 36% of the samples were misclassified as lung adenocarcinoma and 14% as ovarian cancer.

4. Discussion

MicroRNAs expression profiles from original cancer tissues like the ovary and pancreas are difficult to classify because of the heterogeneity of their tissues [14, 23]. However, our classification model classified them correctly, with average accuracy reaching 1 and 0.95 for the ovary and pancreas respectively, using LSTM-FD. Additionally, classification results were 0.97 and 0.93 for ovary and pancreas, respectively, using LSVM-FD. The classification results showed high performance when detection lung cancer that were 1 and 0.96 using LSTM-FD and LSVM-FD respectively. Even though LSTM is designed for sequential data, it could be adapted to detect miRNA expression profiles that suggest there is autocorrelation among miRNAs features. Limited studies applied LSTM with miRNAs expression profiles. Muhamed et al. applied LSTM with mRNAs expression and got an accuracy of 0.97 to detect kidney cancer. Previous research using different methods reported accuracy values ranging from approximately 0.8 to 0.99 [18, 20, 30–35]. This depends on the cancer type and model used as shown in Table 7. The study done by [23] applied the least absolute shrinkage and selection operator (LASSO) algorithm and had got a 0.88 accuracy using 13 cancer types. We used three types of cancers from [23] that have the largest data samples compared to other cancer types. Our models reported high accuracy in the range of 0.94 to 0.96 with 100 features set for all models with and without FD and got 0.97 and 0.96 using LSTM-FD and LSVM-FD respectively, with

Table 7. Brief review of studies using miRNAs profile with ML.

Citation	Cancer Type	Sample Size	Data Type	Classifier(s)	Results
Muhamed Ali et al. [20]	Kidney cancer subtypes (Wilms, KIRC, KIRP, KICH, RT)	TCGA: 1,221 patients	miRNA sequences	NCA 35 features +LSTM (500 + 250 units)	Accuracy: 92.9% Balance: 97%
Tasdelen & Sen [18]	Pre-miRNA vs pseudo-miRNA detection	1,124 sequences (417 pos, 707 neg)	miRNA sequences	CNN + LSTM hybrid	Accuracy: 94.3%; Sens: 93.5%; Spec: 94.8%
Bostanci et al. [30]	Colon cancer diagnosis & prognosis	19,472 RNA 168 samples	mRNA	RF, Knn, LSTM CNN	RF accurcay: 97%; Knn: 96% LSTM: 93%; CNN: 93%
Cheerla & Gevaert [31]	Pan-cancer classification (21 tumor types)	TCGA: ~2,588 samples	miRNA expression (~470 miRNAs)	SVM-RBF	SVM-RBF accuracy: 97.2%; 95% (60 miRNAs)
Sun et al. [32]	Pan-Cancer (27 types)	~10,000+ tumor samples from 27 cancer types ~1,046 miRNAs used as input	miRNA expression	sparse autoencoder and MLP	Accurcay: 96.9%
Taghizadeh et al. [33]	Breast Cancer	900 total (762 tumors + 138 controls)	miRNA expression	Logistic regression (feature selection) + MLP	Balanced accuracy: 86%
Nanayakkara J, et al. [34]	Neuroendocrine Neoplasms (NENs)	378 tissue samples (221 NEN, 114 controls)	miRNA expression	multilayer classifier (MLP-like)	Accuracy: 98%
Cheng et al. [35]	HCC (Liver Cancer) Staging Prediction	348 patients: 248 early-stage, 90 late-stage HCC30 selected miRNAs	miRNA expression	MLP; SVM	Accuracy MLP: 66%; SVM: 73%

50 features set. The limitation of our study was the samples size therefore we optimized model parameters to overcome overfitting. Previous studies have reported the effectiveness of using Multilayer Perceptron (MLP) models with miRNA expression data for cancer classification tasks (Table). However, the objective of this study is not to compare different neural network architectures. Our goal is to show the potential of applying Long Short-Term Memory (LSTM) networks to miRNA expression profiles and evaluate their capability to achieve high classification accuracy.

On the other hand, LSVM is an efficient kernel-based classification method that offers lower computational cost compared to traditional non-linear SVMs. It used linear kernel so it avoids the computational overhead of mapping data into higher-dimensional feature spaces therefore it particularly effective for high-dimensional and limited samples size [22]. Even though it shows low accuracy compare to LSTM when number of features is small. Further results indicate most miRNAs samples of pancreas cancer were classified as lung-adenocarcinoma cancers during unseen test of pancreas and most miRNAs samples of lung-adenocarcinoma cancer were classified as pancreas cancers during unseen test of lung-adenocarcinoma as shown in our results in Figs. 5 and 7 respectively. These results need more investigate about the origin of the lung-adenocarcinoma and pancreas cancers. This may consider as primary original cancer (pancreas) and the metastatic one (lung-adenocarcinoma.

5. Conclusion

In this study, we evaluated the effectiveness of two machine learning techniques, Long Short-Term Memory networks (LSTM), and Least Squares Support Vector Machines (LSVM), with Fisher Discriminant Analysis (FD) to discriminate between cancer types. These methods show potential in classifying cancer-related miRNA expression profile data. The effect of FD was noticed for the low number of features set, whereas the accuracy improved for all models with an increased number of features. The LSTM classifiers with FD consistently outperformed LSVM in terms of accuracy but not in computational complexity. The LSTM is able to capture non-sequential patterns of miRNAs expression profiles. Overall, our model offers a promising approach to detecting cancer.

Conflicts of interest

There is no conflict on interest.

References

1. R. A. Smith, V. Cokkinides, and H. J. Eyre, “American Cancer Society guidelines for the early detection of cancer, 2003,” *CA: A cancer journal for clinicians*, vol. 53, no. 1, pp. 27–43, 2003 Jan.
2. D. Crosby, S. Bhatia, K. M. Brindle, L. M. Coussens, C. Dive, M. Emberton, S. Esener, R. C. Fitzgerald, S. S. Gambhir, P. Kuhn, and T. R. Rebbeck, “Early detection of cancer,” *Science*, vol. 375, no. 6586, p. eaay9040, 2022 Mar 18.
3. S. Connal, J. M. Cameron, A. Sala, P. M. Brennan, D. S. Palmer, J. D. Palmer, H. Perlow, and M. J. Baker, “Liquid biopsies: The future of cancer early detection,” *Journal of translational medicine*, vol. 21, no. 1, p. 118, 2023 Feb 11.
4. Y. Peng and C. M. Croce, “The role of MicroRNAs in human cancer,” *Signal transduction and targeted therapy*, vol. 1, no. 1, pp. 1–9, 2016 Jan 28.
5. G. Di Leva, M. Garofalo, and C. M. Croce, “MicroRNAs in cancer,” *Annual Review of Pathology: Mechanisms of Disease*, vol. 9, no. 1, pp. 287–314, 2014 Jan 24.
6. G. A. Calin and C. M. Croce, “MicroRNA signatures in human cancers,” *Nature reviews cancer*, vol. 6, no. 11, pp. 857–66, 2006 Nov 1.
7. J. Lu, G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, and J. R. Downing, “MicroRNA expression profiles classify human cancers,” Nature, vol. 435, no. 7043, pp. 834–8, 2005 Jun 9.
8. H. Azari, E. Nazari, R. Mohit, A. Asadnia, M. Maftooh, M. Nassiri, S. M. Hassanian, M. Ghayour-Mobarhan, S. Shahidsales, M. Khazaei, and G. A. Ferns, “Machine learning algorithms reveal potential miRNAs biomarkers in gastric cancer,” *Scientific reports*, vol. 13, no. 1, p. 6147.7, 2023 Apr 15.
9. F. Hamidi, N. Gilani, R. A. Belaghi, P. Sarbakhsh, T. Edgünlü, and P. Santaguida, “Exploration of potential miRNA biomarkers and prediction for ovarian cancer using artificial intelligence,” *Frontiers in genetics*, vol. 12, p. 724785, 2021 Nov 25.
10. C. Ji, Y. Wang, Z. Gao, L. Li, J. Ni, and C. Zheng, “A semi-supervised learning method for MiRNA-disease association prediction based on variational autoencoder,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 4, pp. 2049–59, 2021 Mar 18.
11. P. Mestdagh, N. Hartmann, L. Baeriswyl, D. Andreasen, N. Bernard, C. Chen, D. Cheo, P. D’andrade, M. DeMayo, L. Dennis, and S. Derveaux, “Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study,” *Nature methods*, vol. 11, no. 8, pp. 809–15, 2014 Aug.
12. W. K. Shams and Z. Z. Htike, “Oral cancer prediction using gene expression profiling and machine learning,” *Int. J. Appl. Eng. Res.*, vol. 12, no. 15, pp. 4893–8, 2017 Aug.
13. W. Kang, V. L. Kouznetsova, and I. F. Tsigelny, “miRNA in machine-learning-based diagnostics of cancers,” *Cancer Screening and Prevention*, vol. 1, no. 1, pp. 32–8, 2022 Dec 2.
14. F. Hamidi, N. Gilani, R. Arabi Belaghi, H. Yaghoobi, E. Babaei, P. Sarbakhsh, and J. Malakouti, “Identifying potential circulating miRNA biomarkers for the diagnosis and prediction of ovarian cancer using machine-learning approach: application of Boruta,” *Frontiers in digital health*, vol. 5, p. 1187578, 2023 Aug 9.
15. D. A. Pisner and D. M. Schnyer, “Support vector machine,” In *Machine learning*, Academic Press, 2020 Jan 1, pp. 101–121.
16. I. Block, M. Burton, K. P. Sørensen, M. J. Larsen, T. T. Do, M. Bak, S. Cold, M. Thomassen, Q. Tan, and T. A. Kruse,

"Ensemble-based classification using microRNA expression identifies a breast cancer patient subgroup with an ultralow long-term risk of metastases," *Cancer Medicine*, vol. 13, no. 9, p. e7089, 2024 May.
17. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–80, 1997 Nov 15.
18. A. Tasdelen and B. Sen, "A hybrid CNN-LSTM model for pre-miRNA classification," *Scientific reports*, vol. 11, no. 1, p. 14125, 2021 Jul 8.
19. S. Aburass, O. Dorgham, and J. Al Shaqsi, "A hybrid machine learning model for classifying gene mutations in cancer using LSTM, BiLSTM, CNN, GRU, and GloVe," *Systems and Soft Computing*, vol. 6, p. 200110, 2024 Dec 1.
20. A. Muhamed Ali, H. Zhuang, A. Ibrahim, O. Rehman, M. Huang, and A. Wu, "A machine learning approach for the classification of kidney cancer subtypes using miRNA genome data," *Applied Sciences*, vol. 8, no. 12, p. 2422, 2018 Nov 29.
21. Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," 2012 Feb 14, *arXiv preprint arXiv:1202.3725*.
22. R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin, "Liblinear: A library for large linear classification," *The Journal of machine Learning research*, vol. 9, pp. 1871–4, 2008 Jun 1.
23. R. Søkilde, M. Vincent, A. K. Møller, A. Hansen, P. E. Høiby, T. Blondal, B. S. Nielsen, G. Daugaard, S. Møller, and T. Litman, "Efficient identification of miRNAs for classification of tumor origin," *The Journal of Molecular Diagnostics*, vol. 16, no. 1, pp. 106–15, 2014 Jan 1.
24. https://www.ncbi.nlm.nih.gov/gds?linkname=pubmed_gds&from_uid=24211363.
25. C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE transactions on Neural Networks*, vol. 13, no. 2, pp. 415–25, 2002 Mar.
26. F. M. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU," 2023 May 27, *arXiv preprint arXiv:2305.17473*.
27. M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information processing & management*, vol. 45, no. 4, pp. 427–37, 2009 Jul 1.
28. A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media, 2019.
29. J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.
30. E. Bostanci, E. Kocak, M. Unal, M. S. Guzel, K. Acici, and T. Asuroglu, "Machine learning analysis of RNA-seq data for diagnostic and prognostic prediction of colon cancer," *Sensors*, vol. 23, no. 6, p. 3080, 2023 Mar 13.
31. N. Cheerla and O. Gevaert, "MicroRNA based pan-cancer diagnosis and treatment recommendation," *BMC bioinformatics*, vol. 18, pp. 1–1, 2017 Dec.
32. S. L. Sun, B. W. Zhou, S. Z. Liu, Y. H. Xiu, A. Bilal, and H. X. Long, "Prediction of miRNAs and diseases association based on sparse autoencoder and MLP," *Frontiers in Genetics*, vol. 15, p. 1369811, 2024 May 30.
33. E. Taghizadeh, S. Heydarheydari, A. Saberi, S. Jafarpoor-Nesheli, and S. M. Rezaeijo, "Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods," *BMC bioinformatics*, vol. 23, no. 1, p. 410, 2022 Oct 1.
34. J. Nanayakkara, K. Tyryshkin, X. Yang, J. J. Wong, K. Vanderbeck, P. S. Ginter, T. Scognamiglio, Y. T. Chen, N. Panarelli, N. K. Cheung, and F. Dijk, "Characterizing and classifying neuroendocrine neoplasms through microRNA sequencing and data mining," *NAR cancer*, vol. 2, no. 3, p. zcaa009, 2020 Sep.
35. Y. Cheng, R. Ban, W. Liu, H. Wang, S. Li, Z. Yue, G. Zhu, Y. Zhuan, and C. Wang, "MiRNA-409-3p enhances cisplatin-sensitivity of ovarian cancer cells by blocking the autophagy mediated by Fip200," *Oncology research*, 2018 Jan 2.